Isolation and analysis of recombinants from mixed virus infections of poliovirus using next generation sequencing (NGS) and bioinformatics

Fadi Alnaji

A thesis submitted for the degree of Doctor of Philosophy

University of Warwick

School of Life Sciences

May 2016

This thesis is dedicated to the souls of my grandparents Om-Jamal and Abu-Jamal.

List of Tables

List of Figures

List of Equations

Acknowledgments

This work was funded by the Saudi Arabia Government, for whom I am very grateful. I would also like to sincerely thank my supervisor Professor David J. Evans for all the super-informative and constructive discussions, for all the wise pieces of advice and beautiful thoughts. I consider myself extremely lucky to have been supervised by Professor Evans, without whom this project would not be as joyful as it was. I could not have imagined having a better advisor and mentor for my PhD study. I would also like to thank Dr. Jonathan Moore for his help to improve the bioinformatics work of this project and a special thank goes for Professor Andrew Easton, who took charge as an acting supervisor when Professor Evans moved to St. Andrews University.

Thanks also go to the members of the Evan's lab who had helped in creating a unique atmosphere to work in. Very special thanks go to my friends Andrew, Zeenah and Jordan. I will keep remembering our great chats – in the presence of black and strong coffee - which had massively helped me to establish many ideas and strategies.

During my PhD, I was privileged to spend two weeks in the laboratory of Dr. Andrew Macadam at National Institute for Biological Standards and Control (NIBSC). Thanks to Dr. Macadam and the members of his lab who were so helpful during my time there.

A very special thanks goes to my wonderful, tender, great, and supportive Mum and Dad for supporting me spiritually throughout writing this thesis and my life in general. A thank also goes for my two younger brothers 'Hamodeh' and 'Ahmed', to all members of my family for keep supporting me, and a very special thank goes to my friends Nour and Hasan who keep inspiring me with their beautiful thinking.

Finally, I must express my warm gratitude to my gorgeous and exceptional girlfriend Nof, for her continued encouragement. Words cannot express how grateful I am; you supported me during the stressful times while writing, and motivated me to strive towards my goal. From my heart and soul thank you for everything. I will love you for ever.

Declaration

This work was completed at the University of Warwick and at the National Institute for Biological Standards and Control (NIBSC) between July 2012 and May 2016 and has not been submitted for another degree. The work is original and unless otherwise stated in the text, has been completed by the author.

Signed

Date

Summary

RNA virus recombination is a key evolutionary mechanism and a driver of genetic diversity. In recent studies using an in vitro "CRE-REP" assay involving replication-compromised parental genomes, recombination was shown to be a biphasic process involving an initial imprecise crossover event which was followed by a resolution process that resulted in the formation of genome-length recombinants (Lowry K. *et al* 2014). We have extended this study to investigate recombination during dual infection by unmodified parental viruses in the absence of selection.

Recombinants were generated by co-infecting HeLa cells with poliovirus type 1 Mahoney and type 3 Leon for 5-hours, followed by RNA extraction, cDNA synthesis, and PCR amplification. Amplified PCR products of both type 1/3 and type 3/1 recombinants were readily detected, cloned individually and sequenced by Sanger sequencing. Within 25 clones sequenced, 18 unique recombination junctions were detected. To get a comprehensive overview of the range of recombination junctions within the virus population the data produced from next generation sequencing of pooled amplified cDNA from dually infected cells was analysed. A bioinformatics pipeline was developed to specifically detect and quantify recombinants within this population.

Three types of junctions were identified, precise (i.e. at the same position in both genomes) and imprecise, including both insertions (as seen in the Lowry 2014 study) and deletions. In an analysis of the P2 region of the poliovirus genome, we identified several hundred different precise and imprecise junctions. The data analysis suggests that recombination is a random event; no correlation between the nucleotide base composition or RNA structure near the junctions' locations of both donor and recipient viral genomes and the recombination frequency was detected. These studies contribute to our understanding of the molecular mechanism of genetic recombination in RNA viruses and suggest ways in which it might be controlled during the development of novel vaccines with reduced recombination potential.

Abbreviations

General Abbreviation	
Bp(s)	Base Pair(s)
cDNA	Complementary DNA
CO_2	Carbon Dioxide
CPE	Cytopathic Effect
CRE	*Cis-acting replication element*
dH_2O	Distilled water
DI	Defective Interfering
DMEM	Dulbecco's Modified Eagle's Medium
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxyribonucleotide Triphosphate
DTT	Dithiothreitol
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EMEM	Minimum Essential Medium With Earle's Salts
FBS	Foetal Bovine Serum
FLC	Full-length Clone
HeLa	Human Cervical Cancer Cell Line
HI	Heat Inactivated
JC	Jukes-Cantor
LB	Luria-Bertani
MFED	Minimal Free Energy Differences
Mg++	Magnesium
NIBSC	National Institute for Biological Standards and Control
NTR	non-translated region
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reactino
pH	Power of Hydrogen
PS	Penicillin and Streptomycin
RdRp	RNA-dependent RNA polymerase
RF	Reading Frame
RNA	Ribonucleic Acid
RNAse	Ribonuclease

RT-PCR	Reverse Transcription Polymerase Chain Reaction
SNP	Single Nucleotide Polymorphism
Virus and Vaccine abbreviation	
BVDV	Bovine Viral Diarrhoea Virus
cpBVDV	cytopathogenic BVDV
FHV	Flock House Virus
FMDV	Foot-and-Mouth Disease Virus
HCV	Hepatitis C Virus
OPV	Oral Poliovirus Vaccine
PV	Poliovirus
PV1	Poliovirus type 1 (species C)
PV2	Poliovirus type 2 (species C)
PV3	Poliovirus type 3 (species C)
PVR	Poliovirus type 3 receptor (CD115)
VAPP	Vaccine-associated Paralytic Poliovirus
Viral proteins abbreviation	
$2A^{pro}$	Virus Protease
2BC	Precursor protein, interacts with cellular membranes
3AB	Precursor protein, binds to one or more CREs
$3CD^{pro}$	Precursor protein with protease activity, binds to one or more CREs
$3C^{pro}$	Virus protease
$3D^{pol}$	RNA-dependent RNA polymerase (RdRp)
VPg	virion protein genome linked covalently to 5’ NCR of virus genome
Genes Abbreviation	
3B	Encodes protein VPg
3A	Encodes a protein that participates in virus replication
2B	Encodes protein that interacts with cellular proteins, induces vesicles in cells
2C	Encodes protein interacts with cellular proteins, induces vesicles in cells

VP1	Encodes Capsid Protein
VP2	Encodes Capsid Protein
VP3	Encodes Capsid Protein
VP4	Encodes Capsid Protein
Bioinformatics and Data Analysis Abbreviation	
BAM	Binary Alignment/Map format
BLAST	Basic Local Alignment Search Tool
Exp.	Expected
FN	False Negative
FP	False Positive
fw-reaad	NGS forward read
INSDC	International Nucleotides Sequence Database Collaboration (INSDC) databases
MetaSim	A sequencing Simulator for Genomic and Metagenomic
NCBI	National Centre For Biotechnology Information
NGS	Next Generation Sequencing
Obs.	Observed
OS X	Apple operating systems
PE	Paired-ends sequencing technology
ROC	Receiver Operating Characteristic
rv-read	NGS reverse read
SAM	Aligning Output Sequence Alignment/Map
TN	True Negative
TP	True Positive
ViReMa	Virus Recombination Mapper Algorithm
Units Abbreviation	
hr	Hour
kb	Kilobase
mg	Milligram
min	Minute
ml	Millilitre
mM	Millimolar

MOI	Multiplicity Of Infection
ng	Nanogram
nt	Nucleotide
nts	Nucleotides
ºC	Degrees Celsius
PFU	Plaque Forming Unit
U	Unit
v/v	Volume per volume total
w/v	Weight per volume total
µg	Microgram
µl	Microliter
µm	Micrometre

1 Introduction

Positive strand RNA viruses, including important human pathogens like poliovirus (PV) and rhinovirus, are known for their ability to evolve rapidly. They possess error-prone RNA-dependent RNA polymerases, which are responsible for the high mutation rate in these viruses. Coupled with their short replication cycle and high yields, this allows for their rapid adaptation to altered environmental conditions such as host immune responses. Additionally, they have the ability to exchange or acquire large regions of the virus genome through the process of recombination, which results in more extensive variations.

Sequence analysis and phenotypic characterisation of numerous animal and plant viruses have suggested that recombination has been a major driving force in the emergence of many natural strains of RNA viruses. Nonetheless, other than an understanding of the underlying principles, little is known about the recombination mechanism and the cellular and viral factors that influence the process. The difficulty in studying recombination stems from the rarity of this biological event, which renders its isolation, from a mixed population, a challenging process.

Poliovirus, a member of the genus *Enterovirus* of the *Picornaviridae* family, is one of the most thoroughly studied and best understood viruses, and was used as a model system in this study. It was used extensively throughout this project; therefore this literature review will focus majorly on poliovirus. The review briefly discusses classification, the life cycle of the virus, evolution and recombination. It reviews the different aspects of recombination as an evolutionary process, including the mechanisms, the types, the influential factors. Moreover, it discusses the role of recombination as an evolutionary process.

1.1 Poliovirus history and virus characteristics

Poliovirus, the etiologic agent of poliomyelitis, was discovered in 1909 by Landsteiner & Popper who experimentally transmitted the disease from man to ape (Flexner and Amoss, 1919). Since then, poliovirus has been intensively studied to better understand and control its pathogenicity. In 1949, John Enders and his colleagues performed a milestone experiment showing that poliovirus could be

propagated in cultured human cells (Enders et al., 1949), which paved the way for the development of highly effective vaccines against poliovirus and facilitated the study of the molecular mechanisms of the poliovirus life cycle in cell culture. As a result, many basic virology techniques were established using poliovirus, such as virus purification protocols (Bachrach and Schwerdt, 1952), tissue culture system (robbins and enders, 1950), and plaque assay analysis (Dulbecco and Vogt, 1954).

1.1.1 Classification

Poliovirus is a member of the genus *Enterovirus* of the *Picornaviridae*, a large family of small – pico is a very small measurement unit – non-enveloped, positive sense single stranded RNA viruses. The genus *Enterovirus* is divided into 12 species with the poliovirus prototype strains (PV1, PV2, PV3) forming the *Enterovirus C* type species (Table 1-1). The Picornavirus family includes other members such as Rhinovirus (cause of the common cold), Hepatitis virus, Encephalomyocarditis virus (EMCV) and Foot-and-mouth disease virus (FMDV). Nevertheless, its classification is being continuously updated and currently consists of 54 species grouped into 31 genera (*Picornavirus Home*, 2016).

1.1.2 Particle structure

Poliovirus possesses a small icosahedral particle approximately 30nm in diameter (Hogle et al., 1985) that is made up of 60 copies of the heteromeric structural unit. Each unit contains one copy of each of the capsid proteins VP1, VP2, VP3, and VP4. Proteins VP1–3 together form the icosahedral shell of the virion while VP4 is distributed on the inner surface of the particle (Racaniello, 2007) (Figure 1-1). Therefore, the poliovirus capsid is described as a pseudo $T = 3$ structure (T is the triangulation number which describes the triangular face in the icosahedral structure, and its subdivisions into smaller triangles). Each protein of VP1-3 contains a ‘core’ composing of an eight-stranded antiparallel beta barrel with two flanking helices. As the protein backbones of β-barrel are folded in the same way among the proteins VP1-3, they possess the same topology. However, the differences among these proteins can be attributed to the loops that connect the β-strands to the N- and C-terminal segments that extend from the central β-barrel domain, these differences are responsible for receptor binding specificities and antigenic characteristics (Hogle et

al., 1985). The interactions between the β-barrel domains of adjacent proteins stabilise the capsid structure around the RNA. The interaction of the C termini of VP1 and VP3 form the depression or 'canyon', around the 5–fold icosahedral vertices of the capsid; the site of the virus receptor attachment to enter the cell (Olson et al., 1993).

1.1.3 Genome organisation

Enclosed within the capsid is the RNA genome which is ~7.4kb in length (Figure 1-1). At the 5' end of the genome the non-translated region (NTR) of 743 nts length is located, followed by a single open reading frame of 2209 amino acids (Kitamura et al., 1981). The first 88 nucleotides of the 5' end of the genome make up a clover leaf-like secondary structure, which has an important role in organising the cellular and viral proteins involved in positive strand synthesis (Andino et al., 1990). From nucleotide 89 to 124 is a region called 'spacer' and until recently, no special function was assigned to the spacer region. However, it was found that this region is strongly affecting the neurovirulent phenotype of poliovirus (Toyoda et al., 2007). The remaining nucleotides of the 5' NTR until the start codon of the polypeptide constitute the Internal Ribosomal Entry Site (IRES), so-called because of its propensity to initiate the translational process (Jang et al., 1989). As opposed to most eukaryotic mRNAs, the RNA of poliovirus does not contain a m7G cap structural that is essential for translational initiation. Instead, a small viral protein, VPg, is covalently linked to the 5' terminus (Lee et al., 1977). At the 3' end of the genome is a 70 nts non-translated region (NTR) followed by poly A (Schwartz and Farman, 2010) which contains two stem loops that may interact to form a tertiary pseudoknot structure (Pilipenko et al., 1992) (Figure 1-2).

1.1.4 Receptors and Tropism

Receptors are cell-surface molecules that bind to the virus and play a fundamental role in the entry of the virus into the cell. Additionally, they can be determinants of tissue tropism; the preference of a virus to invade and reproduce in a particular cell type. This is reflected by the fact that poliovirus can infect primates and primate cell culture but not mice or mouse cell cultures (Flint et al., 2015, p. 123). This is because mouse produce the poliovirus receptor (PVR) CD155 (The mouse orthologue of

human CD155 has been identified as Tage4, cell-surface protein expressed in rat colon). The absence of the PVR amongst mice suggests that the virus did not emerge from an animal reservoir (Mendelsohn et al., 1989).

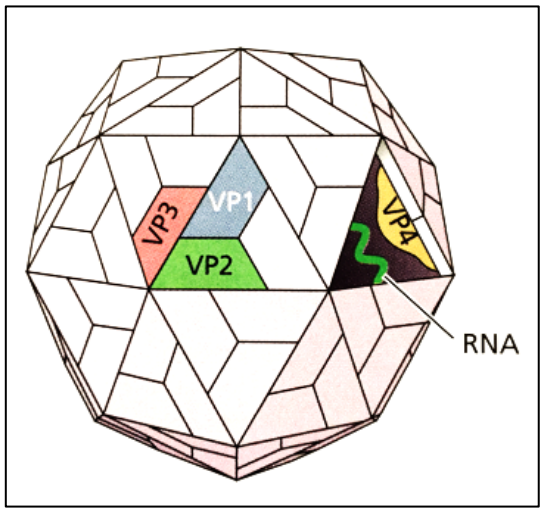


Figure 1-1 Virion structure.
The capsid consists of 60 structural units, each made up of a single copy of VP1, VP2, VP3 and VP4. Examples of these proteins within the capsid are coloured blue, green, red, and yellow. One of the icosahedral faces has been removed to show the packaged RNA strand (green line). This figure is adapted from *Flint et al.* (Flint et al., 2015)

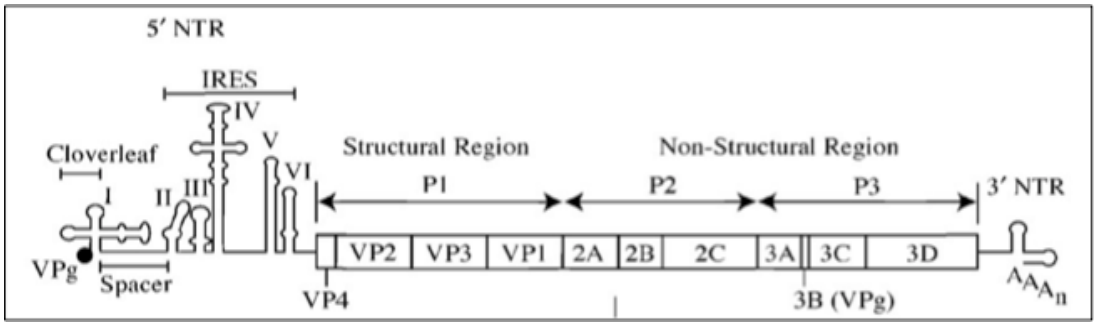


Figure 1-2 Genomic structure of poliovirus type 1 (Mahoney) and proteolytic processing of its polyprotein.
The Poliovirus genome consists of a single-stranded positive RNA that encodes a single polypeptide and is covalently linked to a VPg protein from the 5' end. The cloverleaf and the internal ribosome entry site (IRES) are located at the 5' non-translated region (NTR,). The 3' (NTR) is poly-adenylated. This figure is adapted from De Jesus. (De Jesus, 2007).

Family	Genus	Species	Serotype	Type-Strain	Receptors
Picornaviridae	*Enterovirus*	*Enterovirus C*	PV1	Mahoney	CD155
			PV2	Lansing	
			PV3	Leon	

Table 1-1 Poliovirus strains and classifications
Polioviurs is a member of an *Enterovirus C* species, the genus *Enterovirus* within the Picornaviridae family. All the poliovirus serotypes share the same receptors to enter into the host cell.

1.1.5 Life cycle

Poliovirus enters the cells via attaching to the PVR; once inside the cells the genomic RNA is uncoated and the VPg linked protein is cleaved from the RNA, which allows the translation of the poliovirus genome to proceed (Ambros et al., 1978). The translation product polypeptide is cleaved by virus-encode proteinase $2A^{pro}$ and $3CD^{pro}$ into mature viral proteins (Toyoda et al., 1986, Cameron et al., 2010). The host machinery of the protein synthesis is hindered as a result of cleaving the translation factor eIF4G – that is required for the translation of capped messenger RNA – by $2A^{pro}$ (Svitkin et al., 2001). As soon as enough proteins have been produced, RNA synthesis can begin. The same strand used for protein synthesis is used for minus strand RNA synthesis. Therefore to avoid collisions between the RdRp (moving from 3' to the 5') and the ribosome translating in the opposite direction, there must be a signal for this transition. It is believed that this is mediated by the newly synthesised viral protein 3CD, which together with the poly (rC) binding protein (PCBP), binds to the 5' cloverleaf forming a ribonucleoprotein complex (RNP) to stop the further initiation of viral translation at the 3' end of the viral genome. Another RNP complex is formed by 3CD and the poly (A)-bound protein (PABP) that binds to the poly A tail (Barton et al., 2001, Herold and Andino, 2001).

For the minus strand RNA synthesis to begin (the first step in replication), the two RNP complexes need to interact. It has been suggested that this interaction is a prerequisite for minus RNA synthesis initiation (Herold and Andino, 2001). The second step in replication is the plus strand RNA synthesis, which takes place after several modifications in the partially double-stranded replicative intermediate (RI). The primer for this process is VPg-pUpU , which was found to be synthesised on the

cis-acting element CRE within the *2C* coding region (Goodfellow et al., 2000, Paul et al., 1998). The RNA replication then occurs within a tightly closed rosette-like replication complex composed of vesicles derived from the intracellular membranes (Rust et al., 2001, Egger et al., 2000). After several rounds of RNA replication, the positive strand virus genome is packaged into a newly formed capsid structure by a process called encapsidation (Oh et al., 2009).

1.2 Evolution

Viral evolution indicates the genetic changes, such as mutations, that a virus accumulates during its lifetime, which can arise as an adaptational response to environmental changes or the immune responses of the host. In RNA viruses there are three features that help the virus to evolve rapidly and survive; a) a high mutation rate b) the production of large numbers of progeny, and c) a short replication cycle. Consequently, RNA viruses replicate as a complex and dynamic mutant swarm, called the viral quasispecies (Domingo and Holland, 1997). The quasispecies theory seeks to explain the evolution of life in RNA viruses by exploring the consequences of error-prone replication and large population sizes for genome evolution (Lauring and Andino, 2010). In addition to the high-mutation rate, virus recombination also contributes to the quasispecies nature of RNA virus populations (Freistadt et al., 2007a). This section will discuss the principles and consequences of these mechanisms, focusing on recombination as this project's feature of interest.

1.2.1 High error mutation

All nucleic acid polymerases incorporate incorrect nucleotides during elongation. DNA-dependent DNA polymerases have proofreading capabilities due to the exonuclease activity they possess. These allow them to correct their mistakes. Most RNA-dependent RNA polymerases (RdRp) do not have this capability, which results in high error frequencies in RNA replication (Domingo et al., 2002). The poliovirus RdRp has been shown to have an estimated miss-insertion rate that ranges from 1.2 x 10^{-4} to 1 x 10^{-6} mutation/nt for transition mutations and 3.2 x 10^{-5} to 4.3 x 10^{-7} mutation/nt for transversion mutations (Freistadt et al., 2007a). The low fidelity of RdRp, together with the large population size, contributes vastly to the quasispecies of the RNA viruses. It has been estimated that every possible point mutation and

many double mutations are generated by each replication cycle, and may be present within the population at any time (Lauring and Andino, 2010). The existence of RNA viruses as a mixture of genotypically different viruses makes the selection of viral mutants a relatively frequent process. A good example of this is the selection of the three attenuated poliovirus serotypes that compose the oral poliovirus vaccine (OPV). All the three serotypes were individually serially passaged in primates and primate cells in culture to eventually select variants with different point mutations that are no longer neurovirulent in humans, but can still replicate sufficiently to induce protective immunity. Adversely, it has been demonstrated that a point mutation at position 472 (C472U) of PV3 Sabin OPV within the 5' noncoding region is linked to vaccine-associated paralytic poliomyelitis (VAPP) (Evans et al., 1985). This point mutation was found to reduce the efficience of binding of polypyrimidin tract-binding protein (PTB) to the IRES, which is required for initiation of translation (Guest et al., 2004).

Owing to the high error rates, it has been predicted that RNA viruses with high mutation frequencies exist close to the edge of error catastrophe, *i.e.* viruses may undergo a sharp drop in viability after an increase in mutation frequency. (Crotty et al., 2001, Acevedo et al., 2014). By decreasing the fidelity of RdRp, which in turn will increase the mutation rate, the virus population will be forced into error catastrophe. On the contrary, increasing the fidelity of RdRp will reduce the adaptability of the virus population to the selection pressure in the host (Vignuzzi et al., 2006). However, these observations don't consider the nature of the selection pressures that optimise the evolutionary rate, and therefore cannot fully explain the range of evolutionary dynamics seen in viruses (Duffy et al., 2008). In fact, an extensive sequence analysis has shown that evolutionary rates differed depending on the genomic region studied (Hyypia et al., 1997). Understanding the fidelity and quasispecies theories are crucial in order to make predictions about evolutionary dynamics. Nonetheless, in a biochemical study performed by Fresistadt et al. the poliovirus RdRp fidelity was found to be higher than it should have been as predicted by the poliovirus sequence variation rate. Moreover, they observed a correlation between the mismatches and the RNA dissociation, upon which basis they suggested that the high variations of RNA viruses may be due to recombination (Freistadt et al., 2007a).

1.2.2 Recombination

Genetic recombination is the formation of chimeric molecules by the exchange of nucleotide sequences between genetically related RNA viruses. It is termed recombination when it happens amid non-segmented viruses and reassortment if it occurs amongst segmented viruses (Simon-Loriere and Holmes, 2011). Although RNA recombination and reassortment are mechanistically very different, they both require that two or more viruses infect the same host cell (Egger and Bienz, 2002).

Recombination in poliovirus was discovered in the early 1960s (Agol, 1997). Two different strains of poliovirus that bore mutations providing resistance to either guanidine or horse serum were used to co-infect cells. The yield of putative recombinants (containing double resistance mutations) in the dual infection was found to be 15-20 times higher than the same yield from the single infection (Ledinko, 1963). Using a similar approach recombinants were isolated from the foot-and-mouth disease virus (Pringle, 1965). Subsequently, a recombination map was built by Cooper who used a collection of temperature-sensitive (*ts*) mutants of poliovirus and determined the frequency of recombination between different *ts* markers. Nonetheless, until that time, the presence of double mutants during the mixed infection was the main indicator of the presence of RNA recombination. The definitive proof was provided by King and his colleagues in 1982, who analysed proteins induced by recombinants among the aphthovirus (King et al., 1982, Lai, 1992).

The isolation of a recombinant virus from a natural infection was reported in 1984. After an examination of viruses isolated from children who had received poliovirus vaccines, a recombinant containing sequences derived from the three serotypes of poliovirus was found (Lai, 1992). Subsequently, additional recombinants were isolated from poliovirus vaccinees (Cammack et al., 1988, Cuervo et al., 2001). This indicates that poliovirus recombination can occur naturally and thus might contribute to outbreaks of poliomyelitis in areas of low vaccination coverage (Kew et al., 2002).

1.3 The mechanism of recombination

The mechanism by which recombination occurs can be conceived in two different ways: 1) copy-choice mechanism and 2) break-joining mechanism. The former is coupled with the process of genomic replication, and is widely cited as the most favoured mechanism of recombination, while the latter does not depend on replication occurring.

1.3.1 Copy-choice mechanism

The copy-choice mechanism is a mechanism by which a synthesis of the negative RNA strand is interrupted on the donor strand for the reasons discussed below. As a result, the incomplete nascent strand dissociates – either with the RdRp or without - from the original template and serves as a primer to resume elongation on the acceptor strand. The product of this process is a chimeric negative sense RNA derived from both parental genomes (the donor and the receptor) (Lai, 1992). This mechanism has been commonly accepted since the work of Kirkegaard and Baltimor on recombinants in poliovirus. They co-infected cells with wild-type poliovirus that was guanidine sensitive and temperature resistant with a poliovirus that bore double mutants which made it guanidine resistant and temperature sensitive. After 6 hours, they collected the progeny and looked for recombinants that were both guanidine and temperature resistant (Kirkegaard and Baltimore, 1986). The results of their experiments were strongly in favour of copy-choice mechanism. Additionally, it could be predicted - under restrictive conditions for one or the other of the parental genomes – to differently affect the frequency of recombinants if they were generated during the positive and/or negative strand synthesis. Their experimental results strongly supported the hypothesis that recombination occurs during the negative RNA strand synthesis (Agol, 1997). Moreover, two factors may further support this hypothesis. First, it is suggested that the recombination frequency depends heavily on the availability of the acceptor strand (Jarvis and Kirkegaard, 1992), taking into consideration the fact that positive strands are much more abundant than negative strands in the infected cell. Second, the availability of the negative strands would be lower than the positive strands due to the fact that negative strands are involved in

the replicative intermediate structure. Therefore they are likely to exist predominantly in a double-strand rather than in a single strand (Agol, 1997).

1.3.2 Forced copy-choice and non-replicative recombination

Forced copy-choice predicts that a breakage of the donor strand can promote RdRp switching. It is similar to the copy-choice model, except that it requires a break in the donor to occur (Coffin, 1979). In contrast the non-replicative model suggests that a chimeric RNA molecule can be formed without the involvement of polymerase activity (Gmyl et al., 2003). The non-replicative model was suggested to be mediated by a transesterification (Chetverin et al., 1997) or cellular exonucleases and RNA ligase (Scheel et al., 2013).

There is a little doubt that a non-replicative process occurs, but the evidence is that it is probably much less efficient. One of the few attempts to quantify and compare the recombinant viruses progeny between the replicative and non-replicative was performed by Lowry *et al.* They found that the recombinants produced by non-replicative assay are only 3.8% of the yield produced by the replicative assay (Lowry et al., 2014). Because of this, the following sections will be within the framework of the copy-choice replicative mechanism, unless otherwise mentioned.

1.3.3 Factors influencing recombination

In an attempt to understand the reasons that cause the RdRp to pause, the results of many separate recombination events have been characterised. For example, it has been suggested that A/U-rich or U-rich sequences promote RdRp slippage, resulting in an accidental incorporation of nontemplated nucleotides at the 3' of the growing strand (Nagy and Bujarski, 1997, King, 1988). Other factors believed to play an important role in promoting the RdRp to pause include the ability of the parental genome to form stable hetroduplexes near to or at the junction location (Romanova et al., 1986, Nagy and Bujarski, 1993). Moreover, the presence of internal stable hairpin structures within one or both parental strand(s) was found to correlate with the distribution of the junction location (Dedepsidis et al., 2010). What also does facilitate polymerase switching is the presence of both strands (donor and acceptor) within the same replication complex. In fact, by using a fluorescent *in situ*

hybridisation technique (FISH), Egger *et al.* were able to locate the RNA sequences of both PV serotypes (PV1 and PV2) in HeLa cells co-infected with both viruses. They found a high percentage (>85%) of individual replication complexes containing both viruses (Egger and Bienz, 2002).

1.3.4 Models of RdRp switching templates

Several models have been proposed as to how the polymerase manages to switch templates, and whether it dissociates from the donor template only or from the nascent RNA strand as well. There are two possible pathways by which the switching might happen, which are termed processive and non-processive (Jarvis and Kirkegaard, 1991). In the processive pathway there are two possible scenarios have been proposed to facilitate the switching.

1. It was postulated that the donor and acceptor template might be brought together and fixed by a third assisting sequence. In this scenario the polymerase may switch templates through shunting rather than dissociation (Kuge et al., 1986).
2. The second scenario wherein several nucleotides at the 3' of the nascent strand –within the complex – may dissociate from the donor strand and anneal to the template strand. The template strand in this case is presumed to accommodate itself within the complex (Jarvis and Kirkegaard, 1991).

On the other hand, in the non-processive pathway, another two possibilities have been suggested to explain the switching of the polymerase between templates:

1. The polymerase dissociates from the donor but not from the nascent strand (Cascone et al., 1990) .
2. The polymerase dissociates from the donor and the nascent strand. In this scenario, it was proposed that the free incomplete nascent RNA strand could be used again in RNA recombination via a copy-choice mechanism (Makino et al., 1986).

1.3.5 Types of recombination

Based on the nature of the RNAs involved and the sites of crossing over, Lai classified recombination into three types (Lai, 1992):

Type I (homologous recombination), recombination in this type involves two closely related or similar RNA genomes with extensive sequence similarity. Moreover, the switching template occurs at related sites between the two parental genomes *i.e* the resulting recombinant RNA retains the exact sequence and structural organization of the parental genomes. Therefore, the word 'homologous' refers not only to the presence of similarity between the two parents, but also to the occurrence of a crossover at a comparable site between the two parental RNA molecules. Usually, this type is used in the literature to refer to the intratypic recombination where the parental genomes are from the same serotypes (Jarvis and Kirkegaard, 1991). However, the term 'homologous' was used by King to refer to 'precise recombinants' in both intertypic and intratypic recombination (King, 1988) Figure 1-3. Examples of homologous recombination can be found in many recombination studies (Jarvis and Kirkegaard, 1992, Runckel et al., 2013, King, 1988, Kirkegaard and Baltimore, 1986, Freistadt et al., 2007b).

Type II (aberrant homologous recombination), recombination of this type involves two RNA molecules with similar sequences. Unlike Type I, the crossover occurs at unrelated – but usually nearby – sites (Figure 1-3). The resulting recombinant contains duplication, insertion or deletion. An example of this type was found between satellite RNAs of turnip crinkle virus (Cascone et al., 1990).

Type III (Non-homologous or illegitimate recombination), in this type, recombination occurs between parental genomes that don't show similarity - see Figure 1-3. This type may account for gene arrangements that are found in the RNA genomes such as insertion and deletion (see next section). Nonetheless, the term 'non-homologous' is used sometimes in the literature to refer to intertypic recombination. This can be exemplified by defective interference (DI) particles (see 1.3.6).

Lai's classification is widely used to differentiate between the different types of recombination. Nonetheless, Nagy *et al.* argued that using the terms 'homologous' and 'non-homologous' to describe RNA recombination was not straightforward and could be confusing (Nagy and Simon, 1997). The confusion that occurs from using such classical terms can be attributed to three reasons that they discussed.

Firstly, the word 'homologous' in Type I implies that the two parental genomes are similar rather than the sequence within which the recombination occurred. This is problematic because short identical sequences (5-15nt) were found to influence recombination (Nagy and Bujarski, 1995); it is quite possible that these short sequences exist among dissimilar RNAs as well. Secondly, the naming system suggested by Lai does not take into consideration other factors that might affect recombination such as the secondary structure. For example, imprecise recombination between two satellite RNAs that share sequence similarity in the region of recombination fits under Type II of Lai's classification. By contrast, imprecise recombination that occurs between satellite RNA and genomic RNA fits under Type III. Nonetheless, both recombination events were found to depend on the RNA structure in the acceptor strand. Thirdly, the term 'homologous' is not always precise as it implies that the sequences in question have a common ancestry (Pearson, 2013). Based on this, Nagy *et al.* has suggested a new classification system that takes into consideration the structure of intermediates, the recombination end products, and the recombination machinery. Briefly, Class I is called 'similarity essential' in which substantial sequence similarity between the parental genomes is required. Class II is called 'similarity non-essential' in which instead of similarity, other RNA features are required such as the RdRp binding sequence, secondary structure, and hetroduplex formation. Class III is called 'similarity assisted', which combines both Class II and I. (Nagy and Simon, 1997).

In addition to that, the terms homologous and non-homologous in Lai's system don't define the characteristics of the recombination junction with regards to the parental virus genome. Therefore Lowry *et al.* suggested the terms 'precise' and 'imprecise' recombinants (Lowry et al., 2014). Moreover, Jarvis C. *et al.* have suggested using the term 'homologous' to refer to intertypic and intratypic recombination. This is because intertypic recombination occurs between viruses of different serotypes

whose sequences typically differ by 10-15%. They used the term ‘non-homologous’ for those recombination events in which little or no similarity is required (Jarvis and Kirkegaard, 1992).

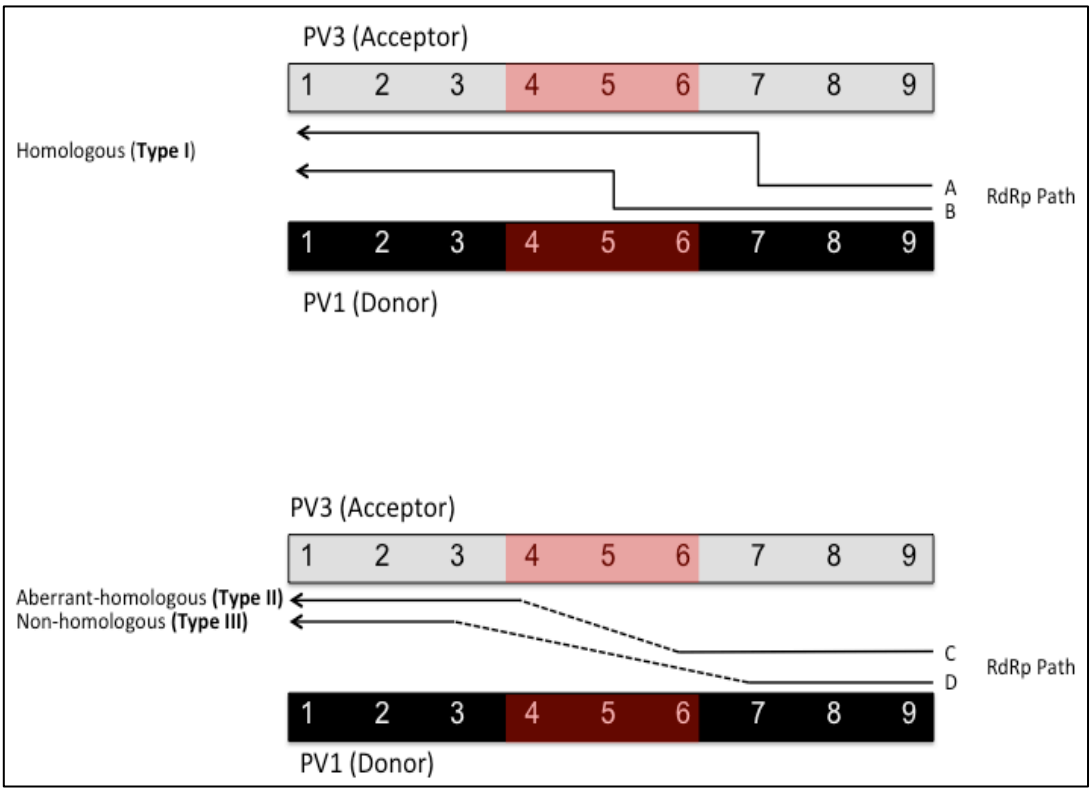


Figure 1-3 Recombination based on Lai’s classification system.
A schematic illustrating the types of recombination as classified by Lai (Lai, 1992). The grey and black rectangles refer to the parental genomes, where their nucleotide composition is depicted by numbers. The arrows between the parents refer to the RdRp paths. The red highlighted regions indicate identical sequences between the parents. The top panel represents Type I homologous recombination, which could happen within a homologous (or non-homologous) sequence (RdRp path A) or identical sequence (RdRp path B). In both cases the resulting recombinant retains the same length and structure of the parental genomes. The bottom panel denotes the two other types. Type II aberrant-homologous is represented by RdRp path C where the switching occurs within an identical sequence and between unrelated nearby sites (the dashed line refers to the part of the genome that RdRp skipped). On the other hand, Type III non-homologous recombination is where the switching involved two unrelated non-homologous sites RdRp path D.

1.3.6 Defective Interfering Particles

Recombinants that have part of their genome deleted can result from a downstream relocation of the RdRp on the acceptor strand during the replication. These recombinants correspond to Type II in Lai's classification if similarity between the parental genomes was involved during the recombination or to Type III if there was no similarity (Lai, 1992). The best-known example of this type of recombination are defective interference (DI) particles. These have lost the capacity to code for all the necessary viral proteins for independent replication and thus are defective in the absence of the parent (also called helper) virus (Pathak and Nagy, 2009). It was first discovered upon serial undiluted passages of poliovirus (Cole et al., 1971). The DI then was reported in many RNA viruses such as the influenza virus (Nayak, 1980) and the Mengovirus (McClure et al., 1980).

In poliovirus, the deletions carried by these DIs map to the P1 region of the genome (largely VP2 and or VP3), and occupy between 4.2 and 13.2 % of the genome size (Omata et al., 1986) keeping the polyprotein reading frame preserved (Agol, 1997). Sequence analysis of the DIs indicated that deletion occurred between nucleotides 1226-2705 (Kuge et al., 1986). The formation of the DI virions was only possible in the presence of viral structural proteins provided by the intact virus (or the helper virus) (Kuge et al., 1986).

RNA viruses (with the exception of the retroviruses) lack the ability to integrate fragments of their genome into eukaryotic host cell chromosomes. Therefore they may rely on alternative mechanisms of attenuation such as DIs to achieve long term virus-host cell association (Perrault, 1981). The generation of DIs is supposed to occur by recombination through copy-choice mechanism (Lundquist et al., 1979) or looped-out mechanisms (Kuge et al., 1986). The latter requires a presence of residues – adjacent to the junction location - that are complementary to other regions in the genome, which may facilitate the formation of transient structures in which the deletion regions are looped out.

It was suggested that the structure of naturally occurring DIs may reflect the mechanism by which they are generated rather than indicating the minimal requirement for their viability (Barclay et al., 1998). However, genome deletions are

restricted to the P1 region since all detected DIs have had their P2 and P3 nonstructural open reading frame maintained (Cole and Baltimore, 1973a). Moreover, engineered out of frame deletions in P1-*2A*-*2B* (in respect to the initiation codon) resulted in a defective genome (Collis et al., 1992). Nonetheless, in the same study, a construct with an in-frame deletion placed at P1-*2A* (nucleotides 910 to 3519) showed the ability to replicate at a significantly reduced level. This indicated that 2A was a *trans*-acting protein and that 2B and possibly other viral proteins were *cis*-acting and showed the importance of viability in determining the structure of the DIs. The DI particles contain all the information needed for replication rendering them potentially useful tools to express foreign genes in place of the capsid proteins (P1 region). This strategy was applied in many studies (Barclay et al., 1998, Lowry et al., 2014, Kaplan and Racaniello, 1988, Percy et al., 1992)

1.3.7 Recombinants contain insertions

Recombinants that contain insertions in their genomes result from an upstream relocation of the RdRp on the acceptor strand during the replication. As in the deletion case, these recombinants correspond to Type II if similarity between the parental genomes was involved, and Type III if there was no similarity (Lai, 1992). Insertion provides important evolutionary opportunities by creating a new genetic materials of novel gene functions (Innan and Kondrashov, 2010). It is important to differentiate between two forms of insertion:

- Form one (gene duplication): insertion that causes gene duplication; a piece of the genome is copied one or more times. There are several previously well-characterised gene duplication events that have shaped extant picornavirus genomes. Examples include the capsid proteins (VP1-3), which share a common structure (Hogle et al., 1985), and the three contiguous VPg proteins of foot and mouth disease (Forss and Schaller, 1982).
- Form two (random insertion): insertion that changes the number of genome nucleotides of a certain gene or region of the genome (insertion), where the source of insertion may be from the same genome or from a cellular genome. This can be exemplified by the insertion of cellular RNA in bovine viral diarrheal virus (BVDV). In tissue culture two, BVDV biotypes,

cytopathogenic BVDV (cpBVDV) and non-cytopathogenic BVDV (non-cpBVDV), can be distinguished (Baker, 1987). Both biotypes are involved in pathogenesis of mucosal disease (MD) in cattle, the most severe clinical manifestation of BVDV (Meyers et al., 1991). The persistent infection with non-cpBVDV is a prerequisite for MD. The cpBVDV can always be isolated from MD animals in addition to the persisting non-cpBVDV (PMC, 1984). When analysing the sequences of the two cpBVDV strains (Osloss and NADL), two different insertions were found, one in each of them. In the first strain, Osloss, the insertion (228 nts) showed a strong homology with the host-cellular ubiquitin gene while the insertion (270 nts) in NADL was almost identical with another bovine mRNA sequence. Based on this, it was suggested that the insertions introduced via recombination between the viral and cellular RNA were responsible for the development of the cytopathogenic viruses (Meyers et al., 1991).

Despite the evolutionary importance of insertion, it has been reported only rarely in RNA viruses, either in natural isolates or cell culture. It has been suggested that one of the main difficulties in recovering recombinants with insertion is due to the length of the genome, which can decrease the fitness of the virus. This because the increase of the genome size is likely to increase the load of mutations and thus reduce fitness (Holmes, 2003). In addition to this, the longer the genome, the more time the replication process would take, which could also be selectively disadvantageous (Simon-Loriere and Holmes, 2013). Nevertheless, Lowry *et al.* have recently used a novel reverse genetic approach in poliovirus that resulted in the isolation of intermediate intertypic recombinants with duplicated sequences at the junction location. The serial passage of these recombinants yielded virus progeny with increased fitness, which had lost the duplicated sequences (Lowry et al., 2014).

1.3.8 Origin and purpose of recombination

Recombination occurs in many RNA viruses and it is believed to play a key role in virus evolution (Worobey and Holmes, 1999). Several evolutionary aspects were found to be associated with recombination such as the expansion of the viral host range (Brown, 1997), the evasion of host immunity (Malim and Emerman, 2001), the evolution of resistance to antivirals (Nora et al., 2007), and an increase in virulence

(Khatchikian et al., 1989). It is a common belief that recombination is a mechanism evolved by viruses to counteract the heavy mutational loads associated with their high mutation rates. That is, it serves to eliminate deleterious mutation and thus contributes to genome conservation and maintaining the replication fitness. Additionally it enables the creation and spread of advantageous traits (Agol, 1997). Although recombination is an important factor in evolution, there is no straightforward answer to the question, why do viruses recombine? The major theories to answer this question can be categorised into two major themes: a) the elimination of deleterious mutation, and b) the creation of advantageous genotypes (Worobey and Holmes, 1999, Egger and Bienz, 2002). Additionally, it has been suggested that the importance of recombination can be determined from the environment within which it happens (Jarvis and Kirkegaard, 1991).

Eliminating deleterious mutation

In a high-mutational population like an RNA virus, the mutation-free genomes become rare and can be lost by genetic drift. This mutational hypothesis is known as Muller's ratchet (Chao, 1990) or the mutational stochastic hypothesis (Hurst and Peck, 1996). Muller observed that the loss is irreversible in asexual populations (which cannot recombine) and that the load of deleterious mutations accumulate in a ratchet-like manner (Muller, 1964). Nonetheless, recombination by combining free mutation parts of each genome can reduce the mutational load in a population (Chao and Tran, 1997). This hypothesis assumes that genomes with a minimum number of deleterious mutations will be lost randomly. Moreover, the effect on the fitness of a new mutation is independent of the number of mutations already in the genome (Hurst and Peck, 1996). This contrasts with the mutational deterministic hypothesis wherein mutations are assumed to synergistically interact (epistasis) so that they stop the destructive effect of Muller's ratchet (Hurst and Peck, 1996). Nevertheless, both hypotheses have been subjected to much debate. For example, are the population sizes of RNA viruses small enough for the Muller's ratchet to take place (Egger and Bienz, 2002)? Or big enough so that the accumulation of deleterious mutations can be counterbalanced by compensatory mutations (Elena et al., 2006) that they can protect the virus from the fatal effects of the deleterious mutations?

Creation of advantageous genotypes

Recombination allows advantageous mutations that arise from different viral RNAs to be united in one genome. It does so by reducing the linkage disequilibrium between mutations, which allows selection to operate independently on individual mutations. This in turn causes the advantageous mutations to accumulate faster (Rice, 2002). However, it is not clear if this alone can explain the purpose of recombination; most cases of antiviral resistance in viruses have been associated with the accumulation of a single point mutation (Egger and Bienz, 2002).

The relative importance of recombination

Jarvis and Kirkegaard have argued that the dynamics of RNA virus populations in different environments could affect the relative importance of recombination for the virus to survive (Jarvis and Kirkegaard, 1991). Based on the assumption that recombination has an evolutionary advantage, they categorised the dynamics within specific environments into three categories, to show the levels of relative importance.

1. Taking into consideration the high error rates of RNA viruses, the larger the population size, the greater the number of mutations that can arise without the need for recombination.
2. In a small founding population subjected to a selective pressure, recombination should play a more important role.
3. This importance is increased even further if more than one allele is involved in creating a specific trait.

1.4 Next Generation sequencing (NGS) and Bioinformatics

Next generation sequencing (NGS) is one of the methodologies that are shaping the understanding of viruses, especially in the area of genome sequencing, evolution, transcriptomics and the discovery of new pathogens. Applying this technology in the field of virology has resulted in an ever growing number of viral genome sequences being deposited within the International Nucleotides Sequence Database Collaboration (INSDC) databases (Brister et al., 2015). As a result of this surge of data, bioinformatics, an analysis tool that uses computer science, has become indispensible for efficiently translating the virus genomes' sequencing data into biological implications.

1.4.1 NGS methods

The studies in virology have largely benefitted from two groundbreaking and revolutionary methodologies. These techniques have allowed virologists to amplify the virus of interest and then sequence the amplicon (Radford et al., 2012), namely Polymerase Chain Reaction (PCR) (Mullis et al., 1986) and Sanger sequencing with chain-termination inhibitors (Sanger et al., 1977). However, the main limitations for these technologies are their restricted scalability and their high cost when applied to a large genome. These limitations, among others, have been overcome through NGS technology, which, based on sequencing of a dense array of DNA, works through iterative cycles of enzymatic manipulation and imaging-based data collection (Mitra and Church, 1999). Although there are several NGS platforms available on the market (454 [Roch], Miseq [Illumina], HiSeq [Illumina], SOLiD [Life Technology]) their work flows are conceptually similar (Shendure and Ji, 2008). Library preparation involves DNA fragmentation, followed by *in vitro* ligation of a common adaptor sequence. Successively, the DNA fragment templates are immobilised on a solid surface or support and then clonally amplified by PCR (Metzker, 2010). The subsequent sequencing process consists of alternating cycles of enzyme-driven biochemistry and imaging-based data acquisition (Shendure and Ji, 2008). Finally, nucleotide sequences obtained from sequencing each cloned fragment of DNA (reads or NGS reads) will be written into output files called FASTQ (Cock et al., 2010). Although most of the NGS platforms share similar biochemical concepts, they use

slightly different methodologies to achieve clonal amplification and sequencing (Barzon et al., 2013).

1.4.2 NGS data analysis (Bioinformatics)

With this phenomenal rate of increase in the data generated by NGS sequencing (Reichhardt, 1999), the application of computational techniques to understand and organise the flux of biological data has resulted in the emergence of novel bioinformatics strategies. The aim of bioinformatics is threefold: to organise data in a accessible way for researchers, to develop tools and resources that aid in analysing the data, and to use these tools to analyse and interpret the data (Luscombe et al., 2001) . Typically, there are two different approaches when using bioinformatics tools to analyse the NGS data. Firstly, if a reference genome is available then the resulting NGS reads can be aligned directly to the reference using an aligner like Bowtie (Langmead et al., 2009). This allows the detection of insertion, deletion, and mutation. Additionally, in the case of recombination where there are two parental reference genomes the NGS reads can be aligned to both; this can be achieved by an algorithm like ViReMa (Routh and Johnson, 2014) (see Chapter 3). Secondly, if the reference sequence is not available, the NGS reads are used to generate contigs (*de novo* assembly); overlapping NGS reads that together represents a consensus region of DNA.

1.4.3 NGS usage to study RNA virus evolution

NGS has been utilised in RNA virus studies to investigate their diversity, evolution and clinical consequences. For example, it was used to study the within-host evolution of hepatitis C virus (Bull et al., 2011) and dengue virus (Parameswaran et al., 2012). These studies and others were carried out mainly to detect the evolutionary aspects other than recombination such as a genetic bottleneck during the infection (Bar et al., 2010), a mutation hot spot (Tapparel et al., 2011), or low frequency escape variants (Henn et al., 2012). Only a few studies have utilised NGS to study recombination in RNA viruses. For example, Routh *et al.* carried out a NGS study to profile RNA recombination within the encapsidated genome of Flock House Virus (FHV) (Routh et al., 2012). Another example is represented by a study performed to establish a high-resolution map of the poliovirus type 1 coding region.

In their study Runckel *et al.* co-infected the cells with a synthetic genome (derived from PV1) containing 368 engineered-markers and wild-type PV1 virus. These synthetic markers allowed them to trace the recombination reads in the NGS dataset (Runckel et al., 2013). The major obstacles in conducting an NGS study on RNA viruses may be attributed to the low frequency at which recombination exists and the possible chimeric artefacts that can be produced by the PCR amplification (Di Giallonardo et al., 2013, Beerenwinkel and Zagordi, 2011). On the other hand, NGS has been employed in identifying recombination events that occur within the genomes of higher organisms. For example, it has been used to detect fusion genes between different chromosomes, an event that has recurrently been implicated in the development of cancer (Maher et al., 2009).

1.5 Aims

Based on this brief review, it is clear that the occurrence of recombination is well established as an important evolutionary key player in RNA viruses. Nonetheless, the underlying mechanisms remain poorly understood. Although NGS provides an efficient approach to studying recombination, the recombination studies using this approach are quite limited. This can be mostly attributed to the paucity of recombination, which complicates the isolation of recombinants from a mixed population *in vitro* and the subsequent detection of recombination NGS reads among the parental reads *in silico.* The aim of this study is two-fold: 1) to develop an approach that can efficiently isolate recombinants, and 2) to establish a bioinformatics method that can be used to detect recombination junctions from the NGS datasets.

Specific aims:

- To develop a reproducible method to specifically isolate and PCR amplify intertypic recombinants between wild-type polioviruses type 1 and 3, in the absence of selection pressures and at an early stage of infection. Note: All the work described in this thesis was performed on the PV1 isolate Mahoney and PV3 isolate Leon. For simplicity these isolates will be hereafter referred to as PV1 and PV3

- To use NGS technology to sequence all possible recombinants generated from the above method.
- To establish a method of using NGS simulation tools to optimise recombination analysis.
- To define a bioinformatics pipeline to detect the junction location from the NGS reads.
- To test the randomness of the recombination event.
- To determine the underlying mechanism of RdRp template switching at the nucleotide level.
- To compare the findings with the original Lowry (Lowry et al., 2014) system to gain an understanding of the recombination event itself and the subsequent resolution process.

2 Methodologies

2.1 Cell culture and virology method

2.1.1 Cell maintenance

Human cervical (HeLa) were grown as monolayers in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 100µg/ml of streptomycin, 100 U/ml of penicillin, 2mM L-glutamine, and 10% heat inactivated (HI) foetal bovine serum (FBS). Trypsinisation was carried out by trypsin-ethylenediaminetetracetic acid.

2.1.2 Virus infection

Cell monolayers were infected with virus at 10 Multiplicity of Infection (MOI). Virus was absorbed onto monolayers for 30 mins at 37ºC/5% carbon dioxide CO_2 air. Virus supernatant was removed and replaced with medium supplemented with 10% H1-FBS. Virus was extracted from the cells 5h post infection before the appearance of cytopathic effect.

2.1.3 Plaque assay

Cells were seeded in 6-well plates and grown to 90% confluency. Ten-fold dilutions of virus stock were made in medium/10% FBS/Penicillin and Streptomycin (PS). Once medium was removed from seeded wells, cells were washed with sterile PBS then inoculated with 500µl of prepared virus and incubated for 30 mins at 37ºC in the presence of 5% CO_2 /air to allow absorption to occur. Plaque overlay media (see section 2.3) was added to each well and plates were incubated inverted for two to three days at 37ºC in the presence of 5% CO_2/air. Cells were stained with crystal violet solution and re-stained post removal of the plaque overlay media. Plaques were counted and the virus titre expressed as plaque forming units per millilitre (PFU/ml).

2.2 Molecular Genetic Techniques

2.2.1 Transformation of *E.coli* with plasmid DNA

Around 3µl ligation mixture (or 1µl plasmid) was added to 50µl ice-thawed α-Select Chemically Competent *Escherichia coli* (Bioline) and stored in an ice bath for 20 mins. The mix was placed in a 42ºC waterbath for 35 seconds before being returned to ice for 2 mins. Five hundred microliters of SOC medium (Sambrook et al., 2000) was added before incubating the tube in a 37ºC shaker (225 rpm) for 1 hour. From this, 100µl (containing the cells) were plated onto Luria-Bertani (LB) agar plates supplemented with appropriate antibiotic selection and incubated inverted overnight at 37ºC.

2.2.2 Plasmid DNA extraction from *E.coli*

Overnight cultures (5 ml or 100 ml) of transformed *E.coli* in LB broth with the appropriate antibiotic for selection were incubated in preparation for mini- and midi-preps respectively. Small-scale isolation was carried out using GeneJET™ Plasmid Mini-prep kit (Fermentas) according to the manufacturer's instructions. Plasmid DNA was eluted in 5µl distilled H_2O (dH_2O). Medium-scale isolations were performed by QIAfilter™ plasmid Midi Kit (Qiagen) according to the protocol provided. The DNA was re-suspended in 200µl dH_2O. A list of plasmids used in this project are provided in Table 2-1.

2.2.3 Virus RNA extraction from cell culture supernatant

QIAamp Viral RNA Mini Kit (Qiagen) was used for RNA extraction from supernatant recovered from the infection according to the manufacturer's protocol. RNA was stored at -80ºC and was thawed in an ice bath when required.

2.2.4 PCR product clean up

Wizard® SV Gel and PCR Clean-Up System was used for purification of DNA from PCR reactions according to the manufacturer's. PCR products were eluted in 30µl and stored at -20ºC until required.

2.2.5 Extraction of DNA from agarose gel

Deoxyribonucleic acid fragments were extracted from agarose gel using DNA extraction kit (Qiagen) according to the manufacturer's instructions. DNA was stored at -20ºC until required.

2.2.6 Ligation of DNA fragments

Digested DNA fragments or PCR products were ligated into vectors at a molar ratio of 3:1 (insert:vector) in a 20µl reaction containing 1x Ligation Buffer (400 mM Tris –HCL, 100 mM $MgCl_2$, 100 mM dithiothreitol (DTT), 5mM adenosine triphosphate (ATP) (pH 7.8 at 25ºC) and 5 U T4 DNA Ligase (Fermentas). Reactions were incubated at room temperature for 2 hrs and used directly within further purification for bacterial transformation.

2.2.7 DNA Sanger sequencing

All sequencing reactions were set up in a 10µl total volume mix using 25pmol of appropriate primer. Plasmid sequencing required 400-500ng, whereas PCR fragments sequencing required 100-400ng. All samples were sequenced by University of Warwick Genomics Facility.

2.2.8 Restriction enzyme digestion

Restriction endonuclease digestion of DNA was carried out using the manufacturer's recommended amount of enzyme in a solution containing 1x of the specific supplied buffer. Incubation temperature and the subsequent thermal inactivation were also carried out according to the manufacturer's instruction. The digestion reaction was incubated from 45 to 60 mins before being run on a 1% weight per volume total (w/v) agarose gel for size/banding pattern analysis.

2.2.9 *In vitro* reverse transcription (cDNA synthesis)

Reverse transcription reactions were carried out using Superscript III reverse Transcriptase (Invitrogen). Ten microliters containing 3µg of purified RNA was incubated in a mixture containing 100pmol oligo dT, 10 mM dNTP mix for 5 mins at

65ºC. Following a 2 mins cool on an ice bath, 0.2µmoles DTT, 1x Superscript Buffer (250 mM Tris-HCL pH 8.3, 375 mM KCL, 15mM $MgCl_2$), and 20 U RiboLock RNase inhibitor (Fermentas) was added and incubated for 2 mins at 42ºC. Two hundred units of Superscript III were added to the reaction before the final 50 mins incubation at 50ºC and reaction termination at 70ºC for 15 mins. The cDNA mixture was stored at -20ºC until required.

2.2.10 *In vitro* transcription

Linearised plasmid for RNA transcription was first prepared by column purification using QIAquickPCR® Purification Kit (Qiagen) and re-suspended in RNase-free water. T7 MEGAscript kit® (Ambion) was used for *in vitro* transcription following the manufacturers protocol. Between 1 and 2µl of RNA was confirmed on a 1% agarose gel before column purification using RNeasy mini kit (Qiagen) and quantified on a spectrophotometer.

2.2.11 PCR amplification

Promega PCR Master Mix 2x was used. A 50µl reaction mix was prepared containing final concentration of the following: 25µl PCR Master Mix 1x (25 units/ml of *Taq* DNA polymerase, 200µM dNTP, and 1.5mM $mgCl_2$), 0.5µm of the relevant forward and reverse primers. For the template, 5µl from the cDNA synthesis reaction was used. Nuclease-free water was added to adjust the reaction at 50µl.

The optimised conditions of the thermal cycling for the two-targeted regions studied in this project were set as follows:

To amplify the P2 region from the end VP1 coding region to the CRE

Denaturation	2 mins at 95°C	
Denaturation	1 min at 95°C	39 cycles
Annealing	30 sec at 58.1°C	
Elongation	1.2 mins at 72°C	
Final Elongation	5 mins at 72°C	

To amplify the P2-P3 region from the CRE to the 3C coding region

Denaturation	2 mins at 95°C	
Denaturation	1 min at 95°C	39 cycles
Annealing	30 sec at 59°C	
Elongation	1.2 mins at 72°C	
Final Elongation	5 mins at 72°C	

2.3 Stock Solutions and Buffers

Crystal violet

0.5g crystal violet powder in 20 ml 100% ethanol, 880ml dH_2O containing 0.9 g NaCl and 100 ml 40% formaldehyde.

DNA agarose gel loading buffer (6x)

25mg bromophenol blue to 3ml 100% glycerol. Make up to 10ml with dH_2O. Stored at room temperature

Plaque overlay medium

10% v/v Minimum Essential Medium (EMEM) with Earle's salts (10x), 1% v/v/ L-glutamin, 3% v/v 7.5% sodium bicarbonate, 2% v/v FCS, 1% v/v PS, and 30% v/v 2% agar.

Agar for plaque overlay medium

2% w/v bacto-agar (Dibco) in distilled water. Microwaved to dissolve powder and stored at room temperature.

2.4 NGS library preparation

Nextera XT DNA kit (Illumina, 2015a) was used to prepare the NGS samples for sequencing. All the steps were carried out according to the manufacturer's protocols. The samples were sequence at National Institute for Biological Standards and Control (NIBSC) on Illumina MiSeq instrument. The generated FASTQ files (Cock et al., 2010), which contained all the sequences and their related quality score information, were uploaded to a specific account on Illumina Basespace cloud (basespace.illumina.com). The files were downloaded later for analysis.

2.5 Recombinant nomenclature

Based on the lack of consistency of the names that have been used to refer to recombinants (see section 1.3.5), it was important in this project to find a suitable way to refer to the recombination events. As this project did not aim to suggest an alternative naming system a simple naming system based on Lowry *et al.* was used that merely served the purpose of this research (Lowry et al., 2014). The recombination that occurred at related site with no deletion or insertion was called 'precise recombination'. If, however, deletion is involved the name became 'imprecise-deletion recombinants', and 'imprecise-insertion recombinants' if the resulted recombinants included an inserted fragment. This naming method does not take into consideration the characteristics of recombinants such as sequence identity or RNA structure. Accordingly these were contextualised based on the type of analysis within the text.

2.6 Statistics used in data analysis

Mann-Whitney U test was used to test the null hypothesis (H_0) that two samples came from the same distribution against an alternative hypothesis (H_a), this particularly was used when tested the randomness of recombination. Spearman test was used to test for correlation between the presence of sequence similarities or

RNA secondary structure and the occurrence of recombination. To measure the significance in difference between the expected and the observed values in the sequence composition analysis Chi-square and Binomial tests were used.

2.7 Computers and clusters

The majority of the bioinformatics analysis (unless otherwise mentioned) was run on the Nero cluster (University of Warwick) through command-line via Terminal on Mac OS X system.

2.8 Bioinformatics algorithms and scripts

2.8.1 Simulation

MetaSim (Richter et al., 2008) was used as the major simulator in this project to simulate the NGS datasets, which were used to optimise the bioinformatics pipeline (chapter 3). Additionally, a custom Perl script was written to simulate recombinants reads for other type of analysis including the sensitivity of detection recombinants (Appendix 1). In both cases the read sequences were saved into FASTA files (NCBI, 2016). Unlike FASTQ files, the FASTA files do not contain any information about the quality score.

2.8.2 Aligning

The algorithm Bowtie2 (Langmead and Salzberg, 2012) was used to map the simulated and the real NGS reads to the reference sequences. The output files were kept for subsequent analysis.

2.8.3 Manipulate the aligning output Sequence Alignment/Map (SAM) file

The resulting SAM file - from the alignment - that contained all the information of the aligning process was further converted into another file called Binary Alignment/Map format (BAM) for some analysis such as the coverage. This was carried out by using the algorithm SAMtools (Li et al., 2009).

2.8.4 Coverage

The coverage of each recombinant, during the optimisation process, was measured by the algorithm BEDtools (Quinlan and Hall, 2010).

2.8.5 Mutation simulation

An online Python script (Appendix 2) was used to insert random mutations into some datasets. The code inserts mutations (the mutation rate is defined by the user) randomly and directly into the FASTA files that contain the NGS simulated reads. The mutation rate was confirmed by calculating the mutation rate of the dataset manually (see next section). On the other hand the Jukes-Cantor (JC) mutation model (see section 3.4.3) was inserted into the dataset by the simulator algorithm MetaSim during the simulation process.

2.8.6 Calculation of the mutation rate in the NGS dataset

The calculation of the mutation rate within the resulted dataset, either with the random mutation inserted by the Python script (Appendix 2) or with the JC model, was carried out in three steps. Firstly, the dataset in-question was aligned by Bowtie2 against its corresponding reference sequence. To ensure that all the reads in the datasets would be mapped regardless of the number of mutations (mismatches) per read, a large value was assigned to the minimum alignment score needed for an alignment to be considered "valid" (*Bowtie 2: Manual*, 2016). This would make the alignment process inclusive so that every reads would map to the reference. As a result, the number of mismatches in each read would be written into the output SAM file. Secondly, the number of mismatches in every read was extracted from SAM file. The SAM file composed of 19 fields separated by tabs, each filed provides information about the mapping process such as the names of the reads, the names of the references, and mapping quality. Field 14, where the number of mismatches was written, was extracted by the bash command ‘cut –f ‘. Thirdly, the number of mismatches was divided by the number of the total bases of the dataset that was generated by MetaSim.

2.8.7 Recombination detection

To detect the recombination junctions from the NGS datasets, the Python-implemented algorithm Virus Recombination Mapper (ViReMa) was used (Routh and Johnson, 2014).

2.8.8 Recombination event extraction from ViReMa outputs

Analysing the NGS dataset by ViReMa produced several output files; one of these would contain the junction location's information. To extract this information, a custom Perl script was written and used for this purpose (Appendix 3). The code was designed with the help of Dr. Jonathan Moore to recognise the format by which the output file from ViReMa was produced. Subsequently, the code would transfer all the information into another text file, which was kept for further analysis.

2.8.9 Recombination read extraction

To confirm the junction location found in the previous steps, some recombination reads were extracted and aligned separately to a reference sequence. To extract the reads that represent the in-question junction's location two steps were considered. First, as ViReMa reported the names of the reads and not the sequences, the names - based on the junction's location - of the reads were extracted by using the bash command 'grep' on Terminal OS X software. The result was written into another text file, which in turn served as an index for the reads' names to extract their corresponding sequences from the major FASTQ file, or FASTA file. In the case of the former, Perl code (Appendix 4) was used to extract the sequences based on the reads' names while for the latter an online small algorithm called faSomeRecords was used. The algorithm faSomeRecords was downloaded from UCSC Genome Bioinformatics Site (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/)

2.8.10 Dotplot

The dotplot graph was generated by Java Dot Plot Alignments (JDotter) software; a Java application with a graphical user interface. (http://athena.bioc.uvic.ca/virology-ca-tools/jdotter/)

2.8.11 Analysis of recombination junction

The sequence composition around the recombination junctions was analysed by several custom Perl scripts (Appendices 5-9), each of which was developed to analyse a different feature. For example, in the case of GC-content, the Perl script would calculate the ratio of GC and AT within a window of 50 nucleotides upstream and downstream to the junction location. In the case of mononucleotides, the ratio of 'A', 'T', 'C', and 'G' would be calculated. To allocate the junctions' locations, the scripts depended on an external tab-separated value (.tsv) file that contained the locations of the junctions on both viruses.

2.8.12 Identical sequence

The lengths of the identical sequences around the junctions were calculated by ViReMa.

2.8.13 Visualisation of the mapped reads

The alignment was visualised by Tablet (Milne et al., 2013), a graphical viewer for visualisation of the assembly and read mappings. Tablet is Java software with a graphical user interface.

2.8.14 Visualisation of the recombinants on the genomes

To visualise recombinants in terms of their locations on both viruses a visualisation method called Parallel Coordinates was used. This was adapted from the original model on the Data Driven Documents website (https://d3js.org).

2.8.15 Text Editing Software

Fraise and TextEdit software were used interchangeably to manipulate the texts. For scripting, the integrated development environment Komodo was used.

2.9 List of the plasmids used in this study

Plasmids' names	Restriction enzyme used for linearisation	Description	Reference
pPV1FLC	*Apa*I	Full-length PV1 Mahoney infectious clone	This project
pT7FLC	*Sal*I	Full-length PV3 Leon infectious clone	(Goodfellow et al., 2000)
pJC105B	*Apa*I	Molecular clone of an imprecise-insertion recombinant	This project

Table 2-1 List of full-length clones used in this study

2.10 The accession numbers of the virus genomes used in this study (NCBI)

Virus	NCBI Accession #
PV1 (Mahoney)	V01149
PV3 (Leon)	K01392

Table 2-2 List of the National Centre for Biotechnology Information (NCBI) accession numbers of the virus genomes used in this project

2.11 List of primers used throughout the study.

Region in the genome	Name	Sequence	Location (5'- 3')	Recombination Products
P2	PV3-3235F	CTCCAAAGTCCGCATTTACA	3235-3255	recP2-R_LD_M
	PV1-4548R	ATCAGGTTGGTTGCTACA	4548-4530	
	PV3-4521R	CAATTAGGTTAGTCGCAACT	4541-4521	recP2-R_MD_L
	PV1-3241F	CTCCAAAATCAGAGTGTATC	3241-3261	
P3	PV3-5647R	GGTGATTTCCAGATTAGTG	5647-5628	recP3-R_LD_M
	PV1-4579F	GCTGAAAGAGAAAACACG	4579-4590	
	PV3-4552F	GCTGAGAAAGAGAACACC	4552-4570	recP3-R_MD_L
	PV1-5610R	GCATCCAAGATCTCCACT	5610-5592	

P2-PV3-Specific	P2-PV1-Specific	P3-PV3-Specific	P3-PV1-Specific

Table 2-3 The primers used throughout the project.
The first column lists the targeted regions within the virus genome. The second column demonstrates the chosen names for each primer. The first part of the name before the hyphen refers to the virus type, at the right of the hyphen is the location regarding the 5' end of the primers followed by direction of the primers (F=Forward, R= Reverse). The third column shows the sequences of each primer. The fourth column lists the location of each primer from the 5' to the 3' end. The last column defines the recombination products that result from each primer pair. The letters 'rec' within the name means 'recombination products', followed by the name of the region within the genome either P2 or P3. At the right of the hyphen the letter 'R' and 'D' refers to the 'recipient genome' and the 'donor genome' respectively, the subscripted 'M' and 'L' refers to the virus strain Mahoney and Leon. For example the name 'recP2-R_LD_M' refers to the recombination product in which the PV1 (Mahoney) is the donor genome and the PV3 (Leon) is the recipient genome. Finally, the colours refer to the type-specific recombinants for each genome. The types and the regions for each colour are demonstrated separately in the bottom raw. Blue = primers specific for poliovirus type 3 within the P2 region. Dark grey = primers specific for poliovirus type 1 within the P2 region. White = primers specific for poliovirus type 3 within the P3 region. Red = primers specific for poliovirus type 1 within the P3 region

3 Bioinformatics (pipeline development)

3.1 Introduction

Viral recombination is a key evolutionary mechanism. Determining whether recombination has occurred and identifying the location of the junctions is therefore of major importance to understanding the underlying molecular process of recombination. With no prior knowledge of recombination junctions, the detection of recombinants becomes complicated. With ever-decreasing costs, the use of NGS has increased significantly in virus discovery and the identification of emerging pathogens. In addition, it offers advantages and new opportunities to study recombinants and recombination. NGS is characterised by the ability to parallel sequence vast numbers of genomic fragments to produce millions of short sequences called 'reads' (see section 1.4.1). Isolating the NGS reads derived from recombination molecules amongst other reads remains a challenge, especially in the context of all the other errors that could be introduced into the NGS dataset through sample preparations and virus replication.

The critical step in identifying the NGS reads derived from recombinants is to map these reads to the reference genome. Current mapping strategies include alignment procedures designed to localize Illumina, Roche, and Ion Torrent reads to a known location in the genome, which can be achieved using software like Bowtie (Langmead *et al.*, 2009), that matches – effectively aligns – sequence reads to a reference sequence. However, whenever an NGS read spans a recombination junction point, part of the read will not map contiguously to the reference, which causes the mapping procedure to fail for that read. Consequently, the read will receive a low mapping score, and eventually end up being excluded from the analysis. For this reason, several approaches have been developed to specifically identify recombination junctions within NGS datasets.

Many recombination detection algorithms are based on the detection of a splice junction in eukaryotic mRNA or in the detection of chromosomal rearrangements or fusions genes in tumours. These include Tophat (Trapnell *et al.*, 2009), Tophat fusion (Kim and Salzberg, 2011), FusionMap (Ge *et al.*, 2011) and MapSplice (Wang *et al.*, 2010). These programmes remove any potential false positive aligned-

read, and putative junctions must be verified by the presence of multiple aligning reads. Filtering the reads based on these two criteria is a valid approach in the case of eukaryotic RNA splicing, as there should be only one or a limited number of biologically correct fusion junctions. Also, these junctions are likely to contain a known splicing site *i.e.* there are sequence motifs characteristic of real splice sites.

However, this is not the scenario when it comes to the viral genomes, which may produce thousands of copies in a single replication cycle. Coupled with the high mutation rate and the short replication cycle, this will create a large pool of viral quasispecie. Therefore, an NGS dataset generated from a viral genome might contain a broad range of possible recombination sites as found in this study and in a previous study by Rrunckel *et al.* (Runckel *et al.,* 2013). Some NGS reads within the dataset would represent different possible recombination junctions, which can be located anywhere in the viral genome and do not necessarily correlate with sequence motifs, which makes the process of recombinants' detection by NGS a challenging process.

An algorithm implemented in Python called ViReMa (Viral Recombination Mapper) was written by Routh *et al.* (Routh and Johnson, 2013) to overcome the difficulty mentioned above. It uses a flexible aligning mechanism; rather than splitting the reads into segments and aligning them independently, it attempts to align the 5' end of a read to the reference genome(s) and then dynamically generates a new read segment. Thereby, generating a new read from the nucleotides at the 3' end of the read that failed to align to the first parental genome. Subsequently, it maps this new read again to the second parental genome. In this way, ViReMa can find a wide range of potential recombinants taking into account the variable nature of the virus population.

In the studies described in this thesis, ViReMa was combined with another algorithm (Bowtie2), and several custom-written Perl scripts to create an optimised bioinformatics analysis pipeline. In the preliminary analysis, many parameters were found to affect the specificity and sensitivity of the pipeline, such as the number of mismatches allowed in the alignment, the location of the junction, and the length of the NGS reads. To model the influence of these parameters, it was decided to generate simulated datasets of known variability and composition and determine the consequences for analysis. Simulated NGS datasets, using parental poliovirus

genomes (type 1 and type 3) and predefined recombinants constructed *in silico*, were generated by the simulation package Metasim (Richter *et al.*, 2008).

3.2 Aims of optimizing the bioinformatics pipeline

a) Identify the junction location of recombinants among NGS datasets.
b) Develop knowledge of the relation between the location of the junction and its representation in the dataset.
c) Evaluate the overall sensitivity and specificity of detecting NGS reads spanning recombination junctions.
d) Assign different sensitivity and specificity profiles for potential recombination junctions based on their location either in the genome or the NGS read.
e) Define a function that can predict the sensitivity of ViReMa detectability over the amplicon.

3.3 Sequencing factors considered for optimising the bioinformatics analysis

There are several factors that can affect the resolution of the output datasets from NGS. They can vary from the sequencing platform to the length of the reads. Based on the nature and the available resources of this project, several factors were chosen to achieve the best quality. These factors were then used in the data simulation and the bioinformatics analysis optimisations.

3.3.1 NGS platform

Despite differences in sequencing methods, all current NGS platforms have several shared features, such as library preparations and amplifications. For this project, Illumina MiSeq was selected to sequence the recombination samples. The main reason is that there are many accessible bioinformatics tools that support Illumina, which are efficient in reducing the analysis time. Additionally, recent studies have shown evidence that MiSeq produces greater depth than other platforms (Frey et al., 2014), which is an important factor when studying rare biological events like recombination in poliovirus.

3.3.2 Targeted sequencing

Targeted sequencing is a sequencing procedure that focuses on a specific gene or a fragment of DNA or cDNA, in contrast to whole genome sequencing (Grada and Weinbrecht, 2013) where the whole genome is sequenced. Sequencing the entire poliovirus genome (7.5kb) with ~300 nts read-length will result in only a small proportion of the reads spanning the recombination junction location. The rest of the reads will match one or other partners in their entirety, thus, will not be informative regarding understanding where the recombination junctions are (Figure 3-1). Based on this, targeted sequencing was chosen, focusing on a ~1.3kb region of the virus genome (nucleotides 3235-4548) within which recombination had already been analysed in a previous study (Lowry et al., 2014).

3.3.3 Number of reads

The number of reads produced by a particular NGS sequencing run determines the depth of coverage of that sequencing run. Increasing the coverage depth was found to increase the resolution of the data (Xie and Tammi, 2009, Medvedev *et al.*, 2010). The combination of targeted sequencing and a high number of reads (coverage) is a common methodology to detect rare Single Nucleotide Polymorphism (SNPs) with high confidence, such as in cancer where the variant can be present at <1% (Schuh *et al.*, 2012). Although identifying SNPs is not the goal of this study, the expected rarity of recombination within a population (Jiang et al., 2007) means that such a combination would be useful to detect and quantify them. Moreover, this procedure will increase the chance of those junctions located at the ends of the amplicon to be detectable in the dataset (see 3.7.1). To achieve this, the number of reads for a 1.3kb amplicon was set to 2-million paired-ends in simulated datasets, to be considered later for the real NGS experiment.

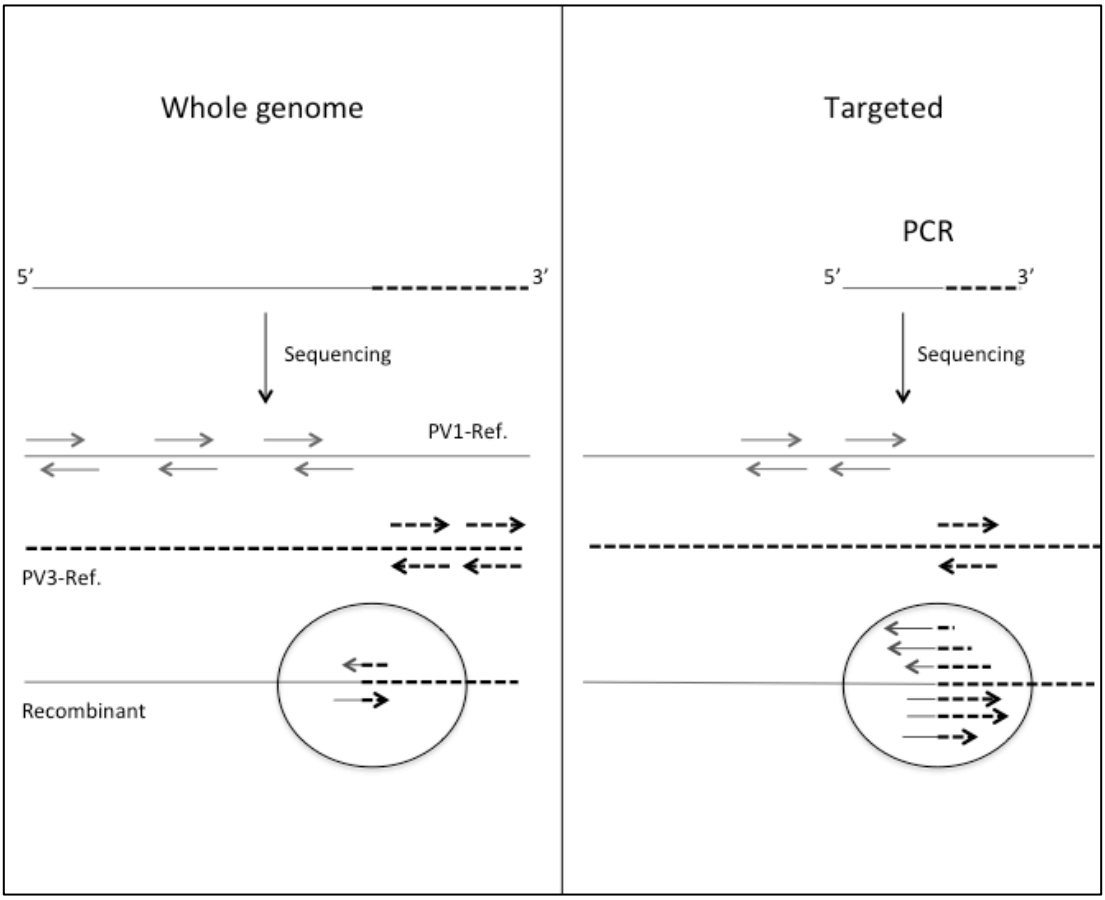


Figure 3-1 Targeted sequencing vs. whole genome sequencing
A comparison between sequencing the whole poliovirus recombinant (left panel) and the targeted region from the recombinant; where the recombination junction occurred (right panel). In the former, no previous treatment was needed while a PCR amplification was needed in the latter. The horizontal two lines represent the recombinants molecule, the grey part derived from PV1 and the black-dashed part derived from PV3. The downward black arrows represent the sequencing step. Reads generated from sequencing are depicted in small arrows. The colour and style of the arrows correspond to the reference sequences they have derived from. The generated reads were first aligned against the parental genome to concentrate the recombination reads in the dataset. The remaining reads represent recombination reads (the reads in the circles), which are aligned to the corresponding recombinant reference

3.3.4 Read-length

The major advantage of elongating the read length is to improve the mappability, also known as the uniqueness (Derrien *et al.*, 2012), of a sequence. The uniqueness stems from the likelihood the longer reads can encompass a unique sequence that anchors all the remaining sequences. This study does not rely on mappability *per se*, as recombinant reads by definition will not align to either parent. Nonetheless, the longer read will provide a higher uniqueness to identify the junction point, increase the probability of a single read spanning a junction site (Töpfer *et al.*, 2013), and increasing the detectable range (see section 3.7.2). The read length chosen was 250-300 nts to optimise the bioinformatics pipeline and to be used later in the NGS experiment.

3.3.5 Fragmentation

In order to sequence numerous molecules in parallel using NGS technology, the DNA has to be fragmented into pieces, this process (termed fragmentation) is indispensable for NGS. It can be achieved either by a hydrodynamic or enzymatic methods. Although the hydrodynamic shearing showed bias towards the linear DNA termini in the fungal genome (Schwartz and Farman, 2010), it was found that both methods perform equally on PCR products, with only minor differences (Knierim *et al.*, 2011). The fragmentation step was performed in the simulation data to produce a ~500nts fragments to best reflect the situation in the real experiment, which was carried out latter by Illumina Nextera kit (see chapter 5) This length was suggested to ensure a uniform sequence coverage depth (Harismendy and Frazer, 2009).

3.3.6 Paired-end reads (PE sequencing)

Paired-ends sequencing technology is the updated version of single-end sequencing. It sequences the same DNA fragment from both ends, thereby facilitating the detection of genomic rearrangements and repetitive sequence elements (Illumina, 2015b). The paired-end reads will increase the sensitivity of finding the true crossovers by giving the algorithm two chances to find the correct junction (Figure 3-2). This is important as ViReMa cannot detect recombination junctions in the last or first 25nts of the NGS reads (see section 3.5). Therefore, the availability of

paired-end reads on the opposing strand should avoid such junctions being missed during analysis. This means the paired-end reads will be used as two different separate reads in this project to inform about recombination, i.e., a FASTA file (see section 2.8.1) of 2-million paired-end reads would become a FASTA file of 4-million single-end reads. All the simulations and sequencing recombinant populations carried out in this research project are Paired-ends reads.

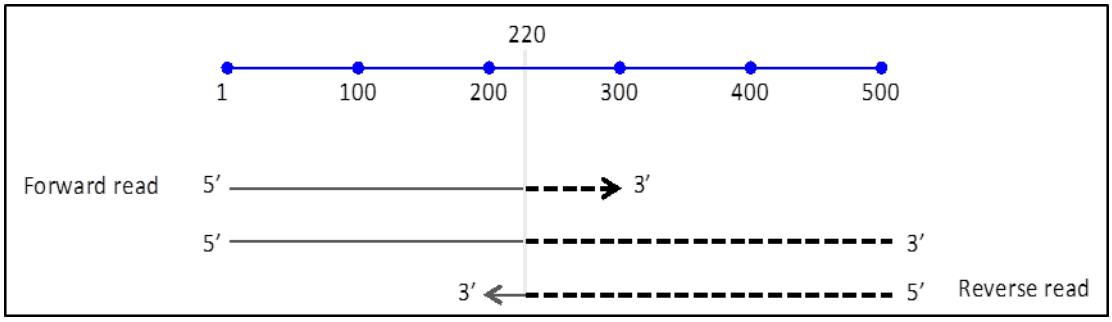


Figure 3-2 Schematic representation of paired-end reads spanning the junction
If the junction is spanned by the last 25 nts of the read, ViReMa would not be able to detect it. However, in paired-end sequencing, the chance for the same junction to be spanned by the opposite read is high. The horizontal grey-black line represents a 500 nts fragment which has resulted from the fragmentation process while the arrows represent the paired-end reads (forward and reverse) with the junction location indicated. The blue line is a measure line marked from 1 to 500 to reflect the length of the fragment and the reads. The transparent vertical line is to facilitate the visualisation of the junction location either in the fragment or the reads. In this example, the forward read spans the junction at position 220, which is a detectable site for ViReMa as it is more than 25 nts away from both ends of the read. On the contrary, the reverse read sees the junction by nucleotide 280, which is undetectable by ViReMa as it located within the last 25 nts of the reverse read.

3.4 Synthetic data.

MetaSim, a simulation algorithm (Richter *et al.*, 2008), was used to simulate a set of NGS datasets to optimize the parameters of the pipeline (Table 3-1). Human poliovirus type 1 and type 3 genomes (see section 2.10) were used as major references, from which, in addition to artificially constructed recombinants, all the datasets were simulated. The simulation maintained all the sequencing factors preferred for this study, besides a range of alternative configurations, which were tested such as mutation rate, and the level of similarities. These were used in various settings to test the sensitivity and accuracy of ViReMa to analyse recombination events.

3.4.1 Synthetic recombinants

Three main recombinants were found in the laboratory by cloning and Sanger sequencing (see chapter 4) were used in the simulated study as major models for optimizations (Figure 3-3). These recombinants were selected on two main criteria. Firstly, each one of them possesses a junction point in a different location, representing three major sites of possible locations for the junction to be placed over the amplicon. Secondly, they show a different level of sequence similarities around the junction location, which could influence the sensitivity of the system. They were named based on the location of the junction with respect to the 5' end of the amplicon. For example 5'-330 is a recombinant with a junction located 330 nts away from the 5' end of the amplicon. A custom Perl code (Appendix 1) was used to generate the recombinant sequences.

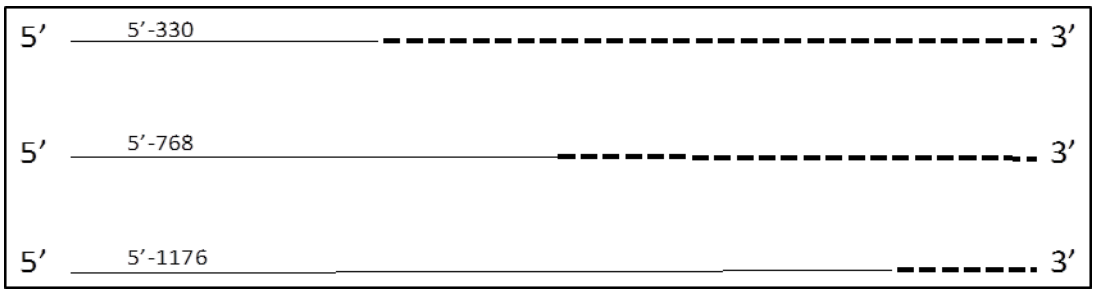


Figure 3-3 Synthetic recombinants reference sequences used in the data simulation
Three main recombinants based on findings from the experimental study were built *in silico*. Each line reflects a recombinant molecule composed of two parts, the thin grey line indicates the PV3 part of the recombinant and the black-dashed line indicates the PV1 component. The numbers above each recombinant indicate the chosen name, which is defined by the distance of the junction from the 5' of the amplicon.

3.4.2 NGS coverage simulation

The theoretical coverage is defined as the number of times that each nucleotide in the sequenced DNA molecule is expected to be sequenced given a certain number of reads of a given length (Sims *et al.*, 2014), and is based on the assumption that reads are randomly distributed across an idealized genome. The depth of coverage for a particular base correlates with the number of reads which see that base. That is, the more each base is sequenced, the more depth of coverage it would receive.

To estimate the number of times a base is expected to be sequenced, the Lander and Waterman (LM) equation was used (Lander, 2015, Lander and Waterman, 1988). This equation (Equation 3-1) measures the depth of coverage by calculating the average number of reads each base in the DNA would receive.

Before simulating all of the datasets, a verification step was carried out to ensure the reliability of Metasim in generating the correct coverage as defined in the configuration. For this purpose, a trial dataset that contained no mutation and was composed of 10000 paired-end reads was simulated to test Metasim (Dataset G, Table 3-1). The NGS reads were first counted in the resulted FASTA using bash command "grep -c" within the terminal *OS X*, which was found to match the defined value passed to Metasim in the configurations. All the reads were then aligned against recombinant 5'-330 reference sequence - which was arbitrarily picked as a source genome for this simulation - resulting in a uniform distribution (Figure 3-4). Coverage was then confirmed by comparing the theoretical coverage which was found to be ~2300 (calculated by the LM equation) with the coverage produced by Metasim (calculated by BEDtools (Quinlan and Hall, 2010)). The coverage was found to be the same as the theoretical coverage, which indicates that Metasim had generated the expected coverage. Note: the words coverage and depth of coverage are used interchangeably in this study, as they refer to the same technical aspect.

$$C = LN/G$$

Equation 3-1 Lander and Waterman equation to calculate the average coverage
C stands for coverage, G for genome length, L for read length and N for the number of reads.

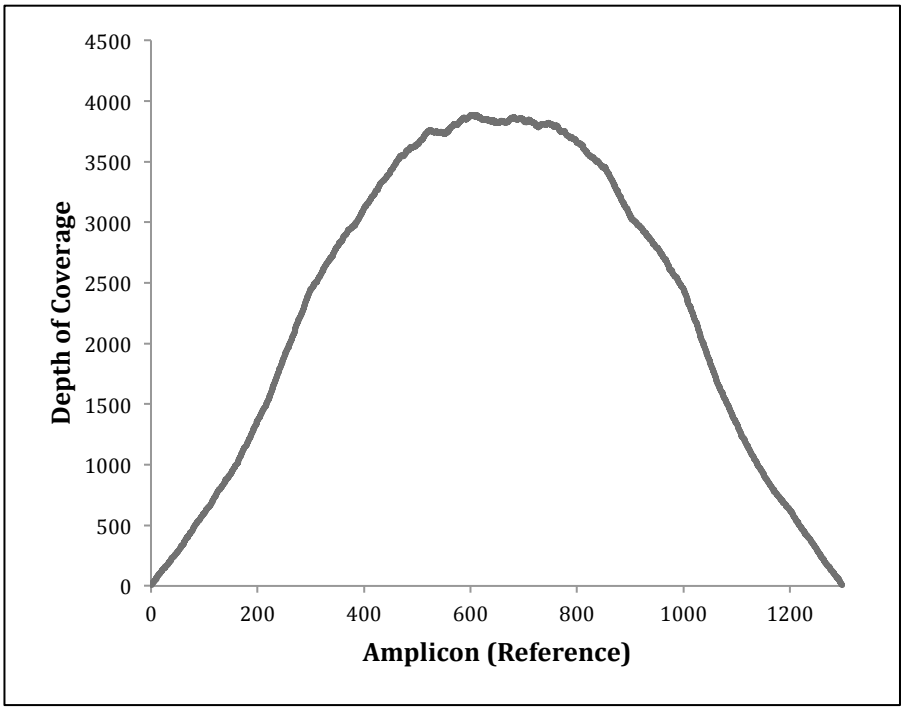


Figure 3-4 Depth of coverage distribution
The coverage calculated from dataset G in Table 3-1. The chart shows a uniform distribution of the reads over the amplicon with uneven coverage coverage. The X-axis indicates the location on the amplicon, and the y-axis refers to the depth of coverage (number of reads).

3.4.3 Mutations & Substitutions in the simulated datasets

Challenges for the interpretation of recombination mapping data are due to the error rate of PCR enzymatic amplification and Illumina sequencing. Moreover, poliovirus exists in a genetically diverse population, as it is liable to error-prone genome replication due to a lack of polymerase proofreading capacity (Domingo, 1992). To optimise the bioinformatics pipeline to work through these errors, simulated datasets with mutations were produced. Two types of errors were introduced to some datasets used in the optimizations (Table 3-1), namely random mutation and substitution. The latter can be divided into substitution by selection and systematic substitution (Figure 3-7).

1) Random Mutation

Mutations result from biochemical processes that can cause genetic changes by introducing errors along the genome (Figure 3-5, left panel). Based on the experimental approach used in this study, several possible sources of such mutations might be considered influential; PCR amplification with an error rate of 10^{-4} -10^{-6} mutations/nt (McInerney et al., 2014), virus replication error with an error rate of $4.5*10^{-4}$ mutations/nt (Wells et al., 2001) and cDNA synthesis process with an error rate of $5*10^{-5}$ mutations/nt (Invitrogen, 2015). The summation of these error rates would bring the total error rate to $5.2*10^{-4}$. This type of mutation was inserted by a script implemented in Python (Appendix 2) which dispersed random mutations within the NGS dataset (Parnell et al., 2011).

2) Substitution

a) Substitution by selection. These are mutations that result from biochemical processes such as replication errors, editing, or nucleic acid damage. However, because their spread and fixation is a population-genetics process that takes place over a much broader scale, the subsequent fixation of mutations in populations is termed substitution (Belshaw et al., 2011, Sanjuán, 2012) (Figure 3-5, right panel). This type of substitution was inserted by applying the Jukes-Cantor model on the simulation process using the default settings (constant rate of 0.01 substitutions). This model assumes

the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C, and G. However, because the experiment developed in this project involved harvesting the virus after 5h post infection, JC model was considered to optimise the bioinformatics pipeline only and does not reflect the experimental condition. For example, JC model would be valid as a reflection of the experimental condition if serial passages were considered.

b) Systematic substitution. Illumina sequencing is a base-by-base sequencing technology using a reversible terminator-based method, enabling the detection of single bases as they are incorporated into growing DNA strands complementary to the template (Huang et al., 2012). Since this technology is reading out one base at a time, the errors introduced are substitutions rather than insertions or deletions. The published substitution rate ranges from 0.1% to 1% , increasing towards the 3' of the reads (Figure 3-6) (Schirmer et al., 2015, Loman et al., 2012, Wright et al., 2011). Illumina substitutions rate were inserted to some of the simulated dataset (Table 3-1).

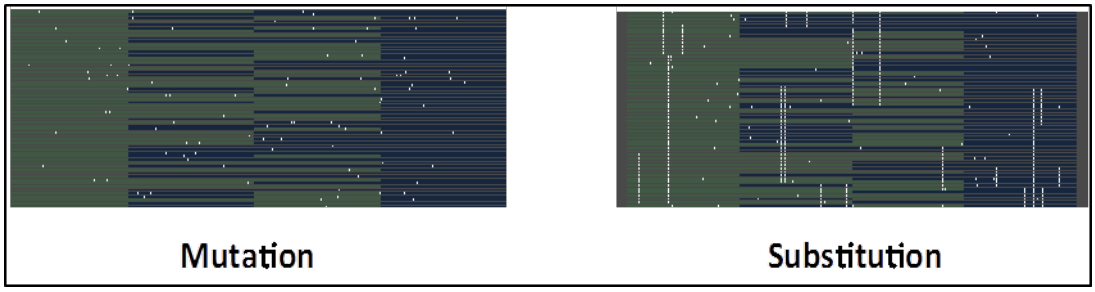


Figure 3-5 Mutation & Substitution simulation visualised by NGS mapping viewer Tablet
Screenshots from Tablet of forward and reverse reads aligned against the corresponding references in Dataset C (left panel) and Dataset D (right panel) (**Table 3-1**). The reads are reflected by four columns separated by different colours; Dark blue = reverse reads, pale green = forward reads. The bright dots in the left panel represent the random mutations in regards to the references, but appear as vertical lines in the right panel when the mutations went through fixation (JC model).

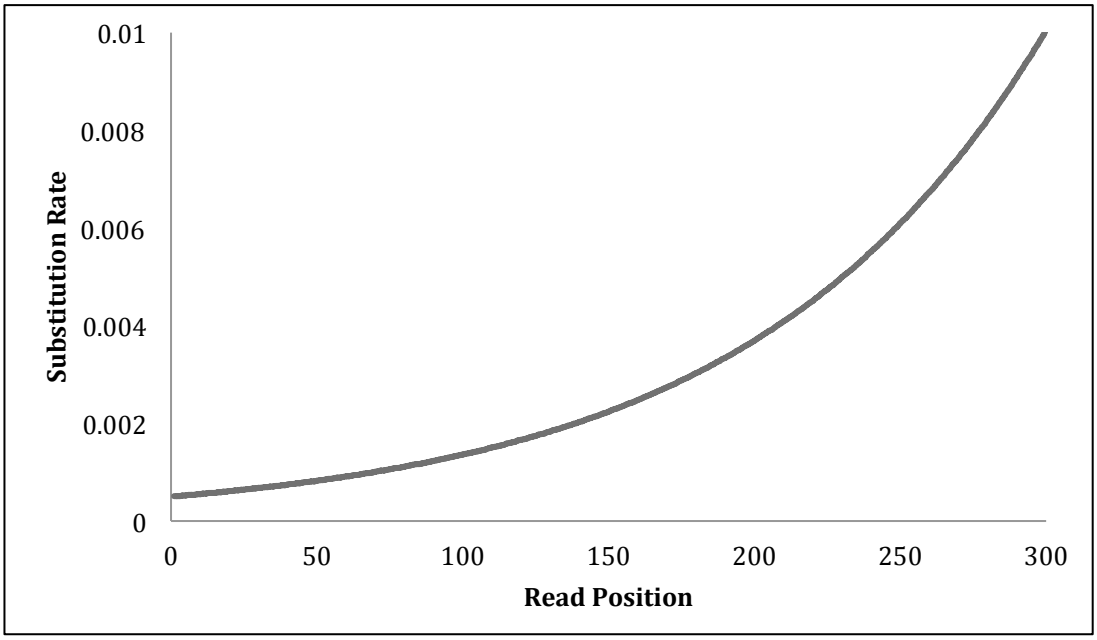


Figure 3-6 Probability curve of substitution errors for a 300 nts Illumina read inserted by MetaSim
The substitution is increasing towards the 3' ends of the reads. The x-axis denotes the position over a 300 nts read. The y-axis represents the substitution rate.

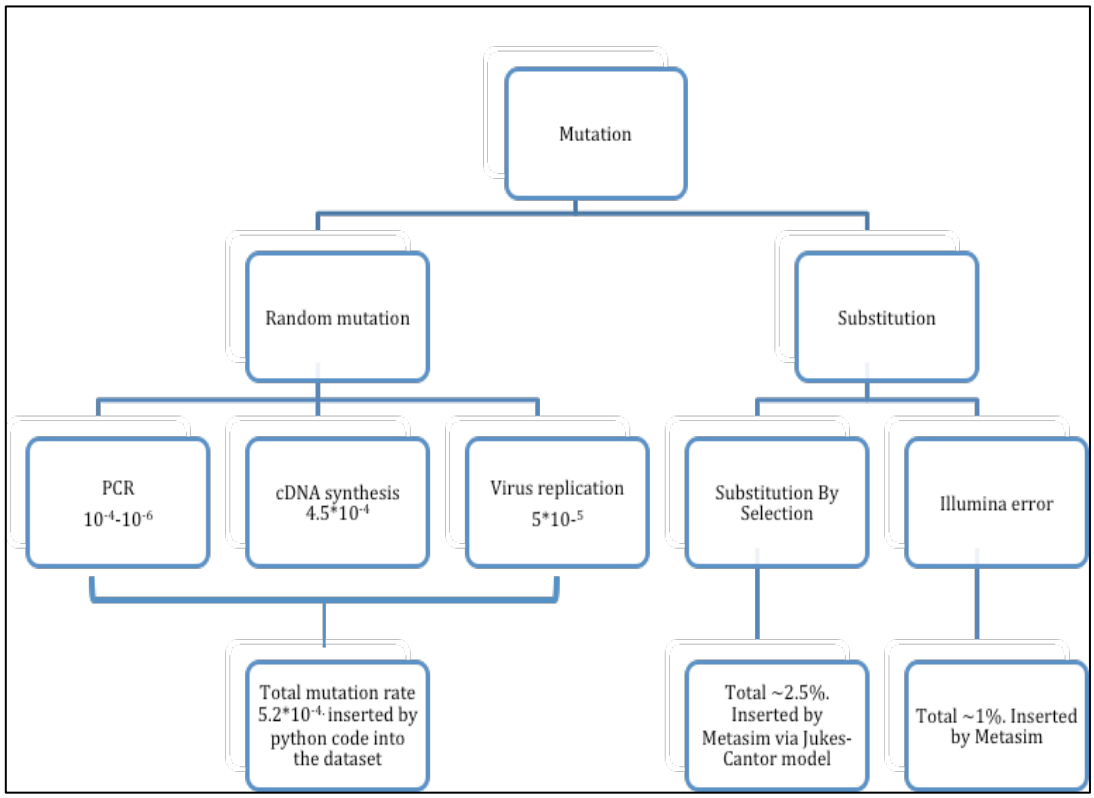


Figure 3-7 Flow diagram of the error sources that inserted into NGS datasets
In the random mutation branch, the rate of every error is written in the same box of a particular error, and the total error of all three is located at the bottom. The substitution branch demonstrates two errors that are likely to occur. These are, Illumina, which was published to be 0.1-1%, and substitution by selection (JC model) which was calculated and found to be ~2.5% from the default settings of the JC model. For the calculation method (see section 2.8.6)

Dataset	Reads number	Variation	Fragmentation length	Source genome sequence	Simulator
A	2 million Paired-End reads	Nil	500 nts (+/- 50 nts)	All the three recombinant sequences: (5'-330), (5'-768), (5'-1176)	Metasim
B		1% Illumina		All the three recombinant sequences: (5'-330), (5'-768), (5'-1176)	Metasim
C		Illumina(1%) + random mutation (0.054%)		All the three recombinant sequences: (5'-330), (5'-768), (5'-1176)	Metasim
D		Illumina (1%) + JC (2.5%)		All the three recombinant sequences: (5'-330), (5'-768), (5'-1176)	Metasim
G	10000 reads	Nil	500 nts (+/- 50 nts)	5'-330	Metasim
I	2 million Paired-End reads	Nil	500 nts	All the three recombinant sequences: (5'-330), (5'-768), (5'-1176)	Metasim
J1	1252 reads (Precise Recombinants)	Nil	Not Applicable	The sequences correspond to the targeted region (nucleotides 3235-4548) in both PV1 and PV3	Rec-generator Perl code
J2	1252 reads (Precise Recombinants)	1% mutation	Not Applicable	The sequences correspond to the targeted region (nucleotides 3235-4548) in both PV1 and PV3	Rec-generator Perl code
K1	1252 reads (Imprecise Recombinants)	Nil	Not Applicable	The sequences correspond to the targeted region (nucleotides 3235-4548) in both PV1 and PV3	Rec-generator Perl code
K2	1252 reads (Imprecise Recombinants)	1% mutation	Not Applicable	The sequences correspond to the targeted region (nucleotides 3235-4548) in both PV1 and PV3	Rec-generator Perl code

Table 3-1 Simulated NGS Datasets used in bioinformatics analysis optimisation

Several databases were created with different features to optimise the bioinformatics pipeline. The ‘dataset’ column on the left shows the names that chosen for each dataset. The ‘reads number’ column includes the number of reads generated for each dataset. The ‘variation’ column presents the type of mutation that was inserted into the datasets and the error rate%. The ‘fragmentation length’ column lists the fragment size of the fragmentation process step in the NGS library simulation. The ‘source genome sequence’ column lists the names of the source genomes used in the simulation to generate the reads. Ref. refers to references in Figure 3-3, PV1 and PV3 indicate complete poliovirus type 1 or 3 genomes. The ‘simulator’ column shows the simulation tool used to generate the data. (Note: When using the Rec-generator Perl code (Appendix 1) as a simulation tool, the fragmentation step was omitted as the code generates reads directly, by appending two sequences together.)

3.5 ViReMa mechanism

ViReMa is a Python script that iteratively calls the small read alignment program Bowtie (Langmead *et al.*, 2009) to try and map all portions of a candidate read. Bowtie aligns the read using a sequence of 20-30 nts length as a seed sequence (hereafter known as seed and set to 25nts throughout this project) and reports the location of a successfully mapped seed. Therefore, ViReMa cannot find the junction's location if it was spanned by the putative seed regions which are the first or last 25 nts of the reads. In this process, Bowtie takes into account the quality scores to report the mismatches, *i.e.* mismatches will be declared as a sum of quality scores if they don't exceed a user-defined value: the "-e-value". ViReMa exploits this method of mapping by purposely assigning a very high '-e' value to allow Bowtie to report a successful mapping of a read regardless of the number of mismatches that follows it. The mismatches are reported in the standard Sequence Alignment Map (SAM) output (Li *et al.*, 2009). From this, ViReMa can decide how many nucleotides from the read have been successfully mapped, and determine the nucleotides that did not map. For example, if a forward read of 300 nts was found to be successfully aligned at its first 150 nts, then this part corresponds to the 5' region of a putative crossover.

The algorithm can be set to allow no mismatch to happen when searching the junction site; as a result, the first encounter mismatch would be considered as a potential recombination junction. ViReMa cuts the read at the mutation site (hereafter known as the cutting-site) and builds another read composed from the mismatched nucleotide to the end of the read. Subsequently, it aligns this subread against the other parental genome before reporting the junction location (Figure 3-8).

Alternatively, the algorithm can allow 1-2 mismatches ('–N' parameter in the command line) to be present in the seed and the remainder of the read. To ensure the differentiation between a mismatch and a real junction point, the mismatch cannot occur among the nucleotides immediately preceding or following a putative recombination site. This process is controlled by the " –X parameter" passed to ViReMa in the command line, which can be set to scrutinize a fixable range of nucleotides before declaring the mismatch a real crossover. For example, if -X set to 5 and –N to 1, ViReMa would allow one mismatch and the cutting-site would be

relocated to the next met mismatch. However, as the –X set to 5, ViReMa checks if there is another mismatch within five nucleotides of the first met mismatch, and if not then the second mismatch will be reported as a potential breakpoint. On the contrary, if a mismatch was found within the five nucleotides, then the first mismatch will disallow the second mismatch, reverting ViReMa back to the first mismatch (Figure 3-9).

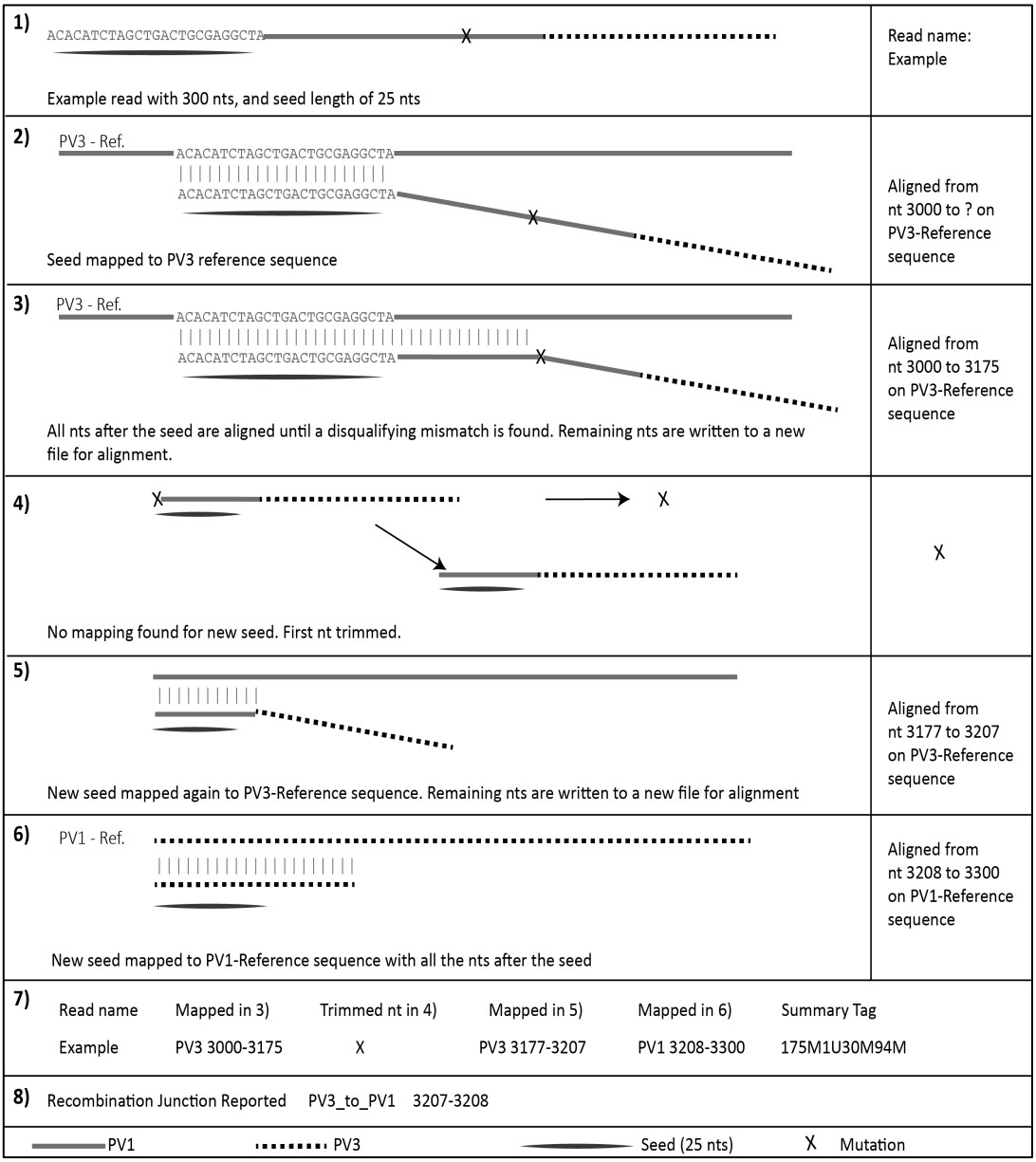


Figure 3-8 ViReMa mechanism with no mutation allowed (N = 0)
The figure illustrates an example read mapped with ViReMa. 1) 300 nts recombinant read with a mutation represented by 'X' in the middle. The first 25 nts of the read represents the seed 2) A good alignment found at nucleotide 3000 in PV3 reference sequence 3) 150 further nucleotides are found to be aligned after the seed until a mismatch is found. The remaining nucleotides thus form a new read. 4) A new seed is extracted from the new subread, but successful mapping cannot be found because of the mutation. Therefore, the first nucleotide is trimmed, and again a new segment is generated. 5) A good alignment is found for the seed and the subsequent 5 nucleotides from 3177 to 3207 until a mismatch is found, this time, the mutation is the junction point. The remaining nucleotides are written in a new file again. 6) Alignment found for the seed and the rest of nucleotides from 3208 to 3300 in PV1. 7) The results of the mapping are recorded in the output file. 8) The name of the recombinant as it appears in the output file. (Note: The length of the seed appears different between the steps, this is for the sake of demonstrating the mechanism in this figure only as they should be in the same length). The figure was adapted from Routh *et al.* (Routh and Johnson, 2014)

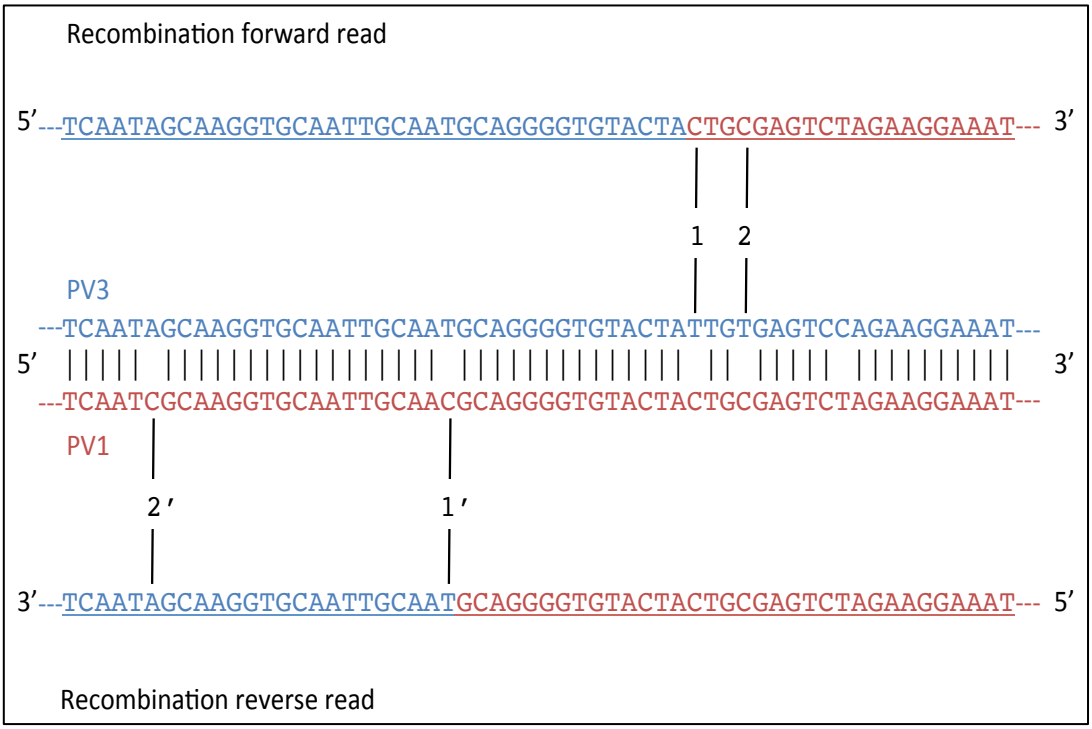


Figure 3-9 ViReMa mechanism in deciding on the junction location when one mismatch is allowed (N=1, X= 5)

The figure illustrates the mechanism by which ViReMa operates when 1 mismatch and 5 similar nucleotides are allowed to happen. The two references are in the middle (PV3/blue and PV1/red). ViReMa will start aligning the 5' side of the recombination read towards the 3 ' end. Above the reference sequences is the forward read of a particular recombinant and lower to the reference sequences is the reverse read. The vertical lines show the mutation sites, and are named 1 and 2 for the forward read and 1' and 2' for the reverse read. Forward read) ViReMa would align the PV3 part first, and when it reached the junction site it would encounter the first mismatch 'site 1 in the figure'. Because ViReMa would allow for one mismatch when N=1, then it would accept this point and move on to the next mismatch 'site 2 in the figure'. At this point ViReMa would check how many nucleotides were aligned correctly between 1 and 2, in this case it is only 2. Therefore the first mismatch '1' would disallow the second mismatch '2', and the location of the first mismatch would be reported as a potential junction location. Reverse read) The lower panel represents the reverse read of the same recombinant. The same strategy is applied on this read. The difference is, however, that ViReMa would find more than 5 aligned nucleotides between 1' and 2'. Based on this, it would mistake the real junction, and rather report 2' as a potential junction point, which can influence the sensitivity and the specificity of ViReMa.

3.6 Measuring ViReMa sensitivity by Receiver Operating Characteristic Curve (ROC)

ViReMa can tolerate up to two mismatches defined by '–N' parameter and allow different lengths of sequence identity between the first mismatch and the putative junction location set by the '–X' parameter. The algorithm can use several combinations of '-X' and '-N' values that can affect the way of searching for recombination junctions and accordingly change the sensitivity and specificity of the analysis. Different combinations of these parameters (termed modes) were tested by Receiver Operating Characteristic Curve (ROC) to define the cut-off values of their specificity and sensitivity in detecting recombination, and based on this the best mode for this study was decided.

A ROC curve graph is a two-dimensional graph in which the true positive rate is plotted on the y-axis (sensitivity) and the false positive rate is plotted on the x-axis (1 - specificity). Since every mode in ViReMa outputs a false positive rate and a true positive rate - together defining a single point in ROC curve space - they are called discrete classifiers (Fawcett, 2006). There are three important points in ROC space that are worth noting. The first one is the lower left point (0,0), which represents the strategy of never producing a positive result; such a classifier commits no false positive errors but also gains no true positives. The second one is the upper right point (1,1), which represents the opposite strategy of unconditionally issuing positive classifications. The third one is point (0,1), which represents perfect classification (100% sensitivity and specificity). The line connects (0,0) and (1,1) meaning no discriminative values; sensitivity = (1 – specificity)

Sensitivity and specificity were calculated by Equation 3-2, wherein three main values were considered:

1) True positive (TP); the number of reads derived from real junctions found by ViReMa.
2) False positive (FP); the number of reads derived from false junctions found by ViReMa
3) False negative (FN); the number of reads derived from real junctions that were not found by ViReMa

Six different modes were considered for ROC curve to test the performance of ViReMa (Table 3-2). These modes would allow ViReMa to tolerate different number of mismatch and different length of sequence identity between the first mismatch and the putative junction's location. ViReMa was used with these six different modes individually to search recombinants in three different datasets (Datasets A, C, and D, Table 3-1), resulting in different sensitivity and specificity profiles reflected by the ROC curve (Figure 3-10).

$$TPR\ (Sensitivity)\ = \frac{TP}{TP + FN}$$

$$FPR\ (Specificity)\ = \frac{TP}{TP + FP}$$

Equation 3-2 Sensitivity and Specificity Calculation
TPR= True Positive Rate. FPR = False Positive Rate. TP= True Positive, FN = False Negative. . FP= False Positive

Mode	Mismatch (-N)	Identical sequence (-X) length
N0X5	No mismatch	5 nts
N0X8	No mismatch	8 nts
N1X5	1 mismatch	5 nts
N1X8	1 mismatch	8 nts
N2X5	2 mismatches	5 nts
N2X8	2 mismatches	8 nts

Table 3-2 ViReMa modes used in ROC curve
The Mode column indicates the values used in each mode; the number next to N defines the number of mismatches allowed, and the number next to X defines the length of the identical sequences that can be tolerated by the algorithm around the junction. The remaining two columns list the values of N and X respectively.

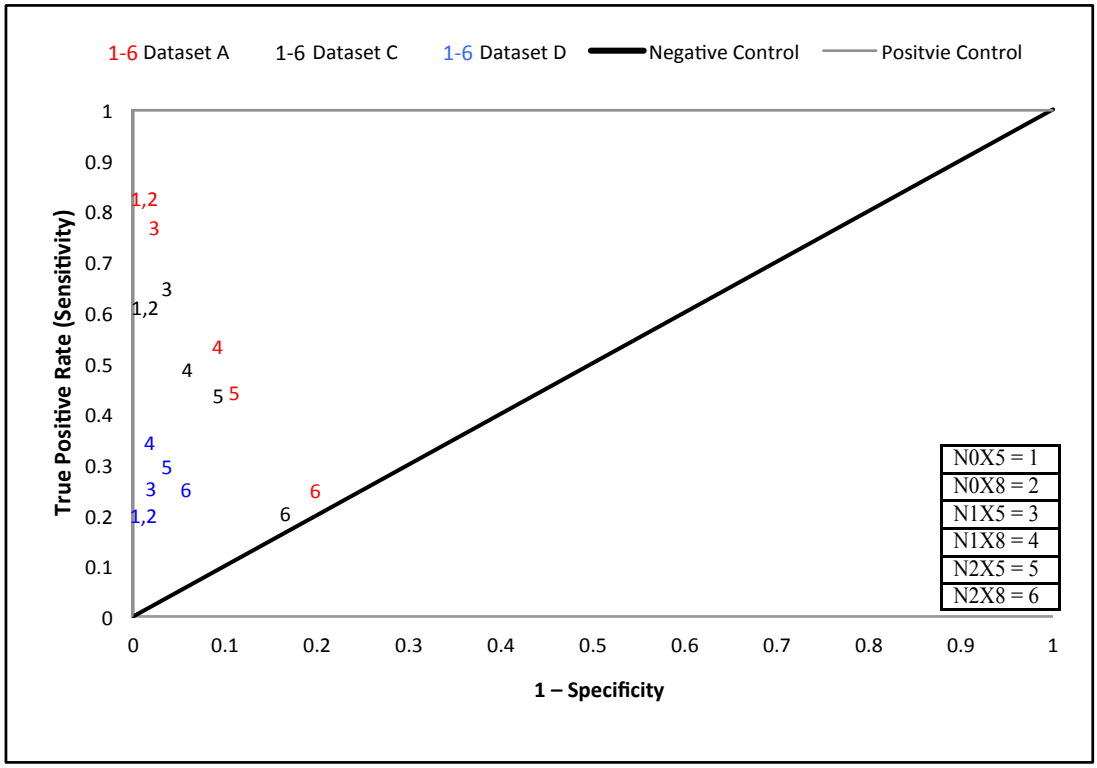


Figure 3-10 Receiver Operating Characteristic Curve (ROC) of ViReMa performance

The ROC curve is built on testing six ViReMa modes using three different datasets with different mutation rates (Dataset A, C, and D from Table 3-1). The table at the right bottom corner assigns a number for every mode to facilitate fitting them in the ROC curve. The x-axis represents the 1-specificity, and the y-axis indicates the sensitivity. The black line is placed to facilitate reading the data, it connects (0,0) and (1,1) (negative control); any classifiers located on this line will be non-discriminative (sensitivity = 1 - specificity). The grey line addresses the positive control; the line passes (grey line) through the point (0,1) reflects the best classifier 100% sensitivity and specificity (note: 100% sensitivity is 1 on the y-axis, 100% specificity is 0 on the x-axis). The numbers assigned to each mode were used to represent the modes in the curve area; they were given different colours to differentiate the datasets. Red, black, and blue denote dataset A, C, and D respectively.

Unsurprisingly ViReMa behaved equally at mode N0X8 and N0X5, because in these modes ViReMa allows no mismatches (-N is zero), thus neither '-X=5' nor '-X=8' were functioning in these cases. This equality is reflected on the ROC curve by placing the numbers assigned for these modes next to each other and separating them by a comma (Figure 3-10) in the curve area. In terms of specificity and sensitivity, these two modes demonstrated the greatest specificity in all of the three datasets analysed, with the sensitivity decreasing as the mutation rate increased. When no errors were permitted in the dataset the sensitivity with modes N0X5 and N0X8 was shown to be ~84% (Figure 3-10, red 1,2), which is the theoretical maximum sensitivity ViReMa could reach with 300 nts read length (see 3.7.2). The sensitivity reduces with the rest of the modes to reach ~20% at the mode N2X8. This drop in sensitivity resulted from the change in the ViReMa mechanism imposed by modes 3-6, which set ViReMa to expect mutation in the datasets. Therefore, ViReMa would report the real junctions as a potential mutation in a dataset that was free of error (Figure 3-10, red 3-6). This was reflected by incorrect positions reported as true junctions (false positive) in the ViReMa output. When ~1% random mutations were introduced to the datasets (Dataset C, black numbers in the curve), the sensitivity generated by modes 1 and 2 decreased to ~64% with ~99.9% specificity (almost no false positives). On the other hand, mode 3 "N1X5" displayed a slightly better sensitivity than modes 1 and 2 with a slight decrease in specificity. The small increase in sensitivity in mode 3 in Dataset C in comparison to Dataset A is explained by the actual presence of mutations, which helped ViReMa to re-find some of the actual recombination junctions. While in Dataset D where a fixed substitution rate of ~2.5% was inserted into the dataset, ViReMa sensitivity and specificity had dramatically decreased (blue numbers in the curve). None of the modes of Dataset D reached any sensitivity higher than ~30% with small differences in the sensitivity and specificity between all the modes, which is reflected by a clustering of the numbers near to each other in the chart.

3.6.1 Erroneous reads that can affect the sensitivity and specificity

ViReMa modes 1 and 2 (hereafter known as N0X5 mode) showed high specificity of ~99.9% and a sensitivity of ~64% with Dataset C (Table 3-1), the dataset that has the closest mutation rate to the estimated mutation in the real data (Figure 3-7). Further investigation was performed on the unreported reads (false negative) and the misreported reads (false positive) which resulted from analysing Dataset C with ViReMa N0X5 mode to determine the nature of these reads. The analysis revealed three types of erroneous reads that ViReMa would either reject or misreport. The first two represent the false negative and the last one represents the false positive.

1. Reads that spanned the junction with their last or first 25 nts (within the seed sequence).
2. Reads that included a mutation within 25 nts of the junction location.
3. Reads that represent junctions located next to an identical sequence and included a mutation right next to the actual junction site.

The first type (constituting 14% of the total) are reads in which the recombination junction was located within the terminal 25 nts of the read, which ViReMa would not detect based on the way the algorithm was implemented (Figure 3-8).

The second type (constituting ~21.9 % of the total reads) are those reads which had a mutation occurring within 25 nts of a junction (Figure 3-11 – read C). Because in N0X5 mode ViReMa would deal with any mutation as a potential junction site; it would cut the read at that mutation site, aligned to the references. If no valid alignment were found, ViReMa would trim the mutation and re-align the read (Figure 3-8). In this type of erroneous read, the presence of the mutation within 25 nts of the junction site would cause ViReMa to cut the read at a site that would generate a subread that had less than 25 nts from one parent and the rest from the other parent. Subsequently, the chance that ViReMa would find a seed for the next round of alignment was rare. For example, if the junction lies at location 150 of a 300 nts read, and the mutation lies at nucleotide 140 of the same read, then ViReMa would cut at nucleotide 140. The resulting 160 nts subread would contain 10 nucleotides from one parent and 150 from the other. In this case ViReMa would not be able to find a unique seed, unless it trimmed 10 nucleotides from this read. In this

process, ViReMa would consider the trimmed 10 nucleotides as an insertion and would not report this read; rather it would keep it in a separate file called ‘unknown recombinants’. However the junction site could be detected if the number of trimmed nucleotides is less than or equals 5. This is the default value and can be modified from the ViReMa settings.

The third type (constituting 0.1% of the total reads) are misreported reads due to a misincorporation that occurred right next to the actual junction locations located next to an identical sequence (Figure 3-11 – read D). For instance if the junction - in a recombinant amplicon composed of PV3 in its first 1000 nts and PV1 in its last 300 nts - locates near to a short identical sequence of two nucleotides, ViReMa would report it correctly if there was no mutation (Figure 3-11 – read A). Nevertheless, if a mutation occurred at nucleotide 1001, the starting part of PV1, and happened to be the same nucleotide as PV3 at that location, then ViReMa would report the junction three nucleotides downstream to the actual junction. As a result, a read with the wrong junction location will be reported.

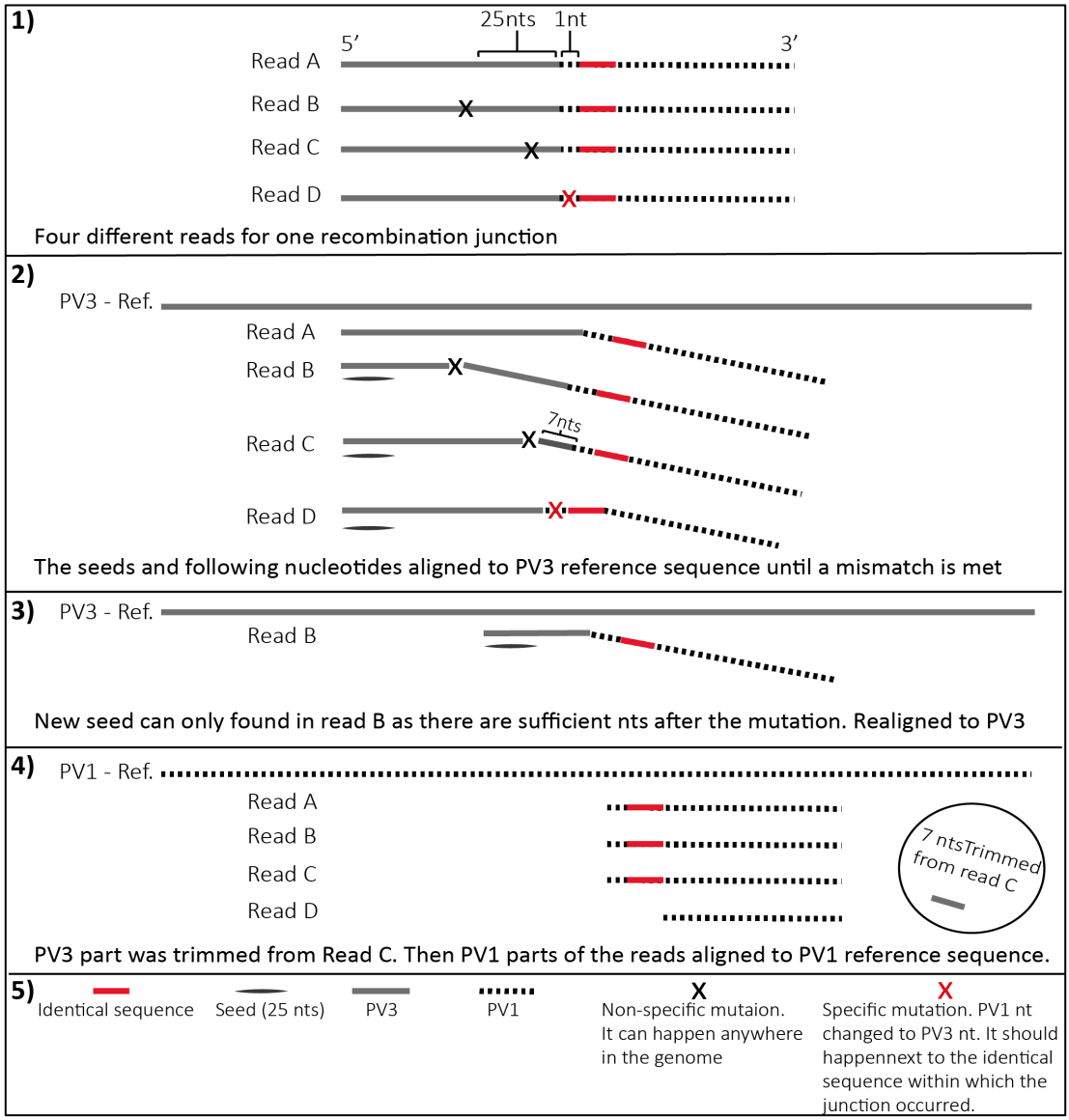


Figure 3-11 The mechanism of ViReMa in reporting false negative and positive recombination junction
The first panel shows four different reads for one recombinant that has its junction located at the connection point between the grey and black-dashed line. Downstream to the junction, there is one nucleotide that belongs to PV1 before an identical sequence is encountered (the red line), then the remaining nucleotides are unique for PV1. The black X refers to a nonspecific mutation, *i.e.*, it could happen anywhere, while the red X is a mutation that occurred at the nucleotide belonging to PV1 before the identical sequence changing it to the same nucleotide that PV3 possesses at this location. By this mutation, the identical sequence has become one nucleotide longer; therefore, it is coloured red (Read D). In the second panel, ViReMa mapped the reads to the PV3 reference until it encountered the mutation. Read A) It is a straightforward recombinant read without any mutation, the first part aligned to PV3 and the other part aligned to PV1 reference (shown in the fourth panel). This read will appear in the result output file with the correct junction location. Read B) ViReMa cut the read at the mutation site, formed another seed, and re-aligned the new seed to PV3 reference (shown in panel 3). The rest of the read aligned successfully to PV1 reference sequence (shown in panel 4). This realigning process happened because ViReMA found enough nucleotides (>25) to form a new seed after the cutting site. This read will appear in the result output file with the correct junction location. Read C) ViReMa cuts at the mutation site, it did not find valid nucleotides to form a new seed; as a result, it started to trim the nucleotides and re-aligned the trimmed read. After trimming the 7 nucleotides of PV3 part of the read and finding a satisfactory alignment for the PV1 part, ViReMa would consider this as an insertion (as it is > 5 nts) and save the read in another file as unknown recombinant. The trimmed part is showing in a circle in the fourth panel. Read D) This read has a mutation on the PV1 nucleotide that was located before the identical sequence. ViReMa would not recognise this mutation, as it happened to be identical to PV3 nucleotide at this location. Rather, it would consider this nucleotide part of PV3 and would report an incorrect junction location located after at the 3' end of the identical sequence. This read would appear in the output results file with an incorrect junction location.

3.7 Different elements affect the sensitivity of ViReMa

As shown in the ROC curve graph (Figure 3-10), the presence of mutations in the dataset would affect the sensitivity and specificity of finding recombination junctions by ViReMa. To cope with such mutations, ViReMa utilises different modes that adjust the algorithm mechanism. Nonetheless, mutation is not the only factor that impacts on the process of searching recombinants' reads in an NGS dataset by ViReMa. In this section, there are three elements – divided into categories – that can influence the searching capacity that was studied. They are either technical or biological factors:

- Technical
  a) Location of the junction in the amplicon
  b) Location of the junction in the NGS reads
- Biological
  a) The context of the junctions: similarity and divergence.

3.7.1 Analysing the effect of the location of the junction in the amplicon on the sensitivity ViReMa

The Lander and Waterman assumption (Lander and Waterman, 1988) is that the NGS reads are randomly distributed across the genome. This is also reflected in the random nature of fragmenting the DNA in the real experiment (Syed *et al.*, 2009). Therefore, the extremities of the amplicon are liable to receive less coverage than the internal region. Consequently, this will result in a lower number of reads representing recombination junctions located at the ends of the amplicon. For example junction A in (Figure 3-12) is covered by 3 reads, while only 1 read covers junction B.

To examine how the NGS reads' underrepresentation of the termini would affect finding recombination junctions, Dataset A (Table 3-1) was used. The reads of this dataset correspond to the three primary recombinants, used in this study (Figure 3-3) with no mutations applied. The reads were first aligned against the three reference sequences (Figure 3-3) using Bowtie 2 (Langmead and Salzberg, 2012), to isolate and quantify the reads that cover each junction. Bowtie 2 was configured to not allow any mismatches to prevent non-specific alignments. The aligned reads were

then separated into forward and reverse reads, BAM files (see section 2.8.3) generated using SAMtools and finally the coverage calculated by BEDtools (Quinlan and Hall, 2010) (Figure 3-13). The average coverage was calculated from the files produced by BEDtools, and charted in (Figure 3-14).

From the total number of reads (4 million reads), the maximum coverage for all the recombination reads was 11.7%. Dissecting the coverage by looking at the coverage for each synthetic recombinant (see section 3.4.1) of each read orientation (forward and reverse), three different patterns can be observed. Recombination junction 5'-330 was found to be predominantly covered by the forward reads (~3%) and partially covered by reverse reads (~0.6%). As for recombination junction 5'-768, it was equally covered by forward and reverse reads (~3%) and recombination junction 5'-1176 was covered by the reverse reads (~1.2%) and hardly covered by the forward reads (0.007%). From these patterns, it is clear that the closer the junction to the ends of the amplicon, the less depth of coverage it receives. How this can be translated in the context of ViReMa is discussed in the next section.

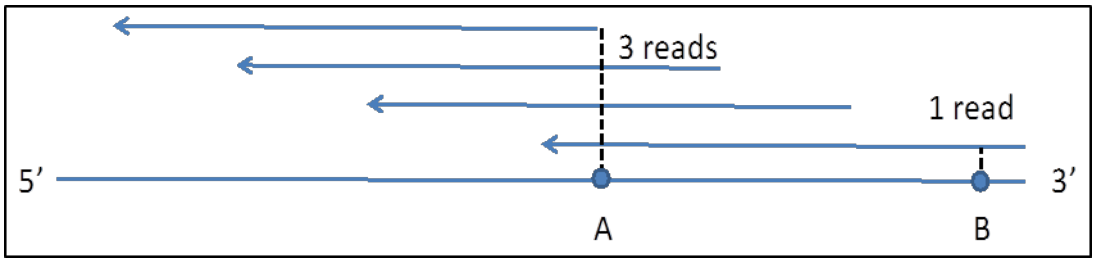


Figure 3-12 Schematic drawing illustrates the level of coverage for recombination junction located close to the end of the amplicon
In this example, the coverage of two different junctions is demonstrated. The circles refer to the junctions' locations. The arrows indicate the reads. When the junction is closer to the end, it is covered by one read (Junction B), but when it is closer to the middle, the number of reads increased to three reads (Junction A). The bottom blue line denotes the amplicon.

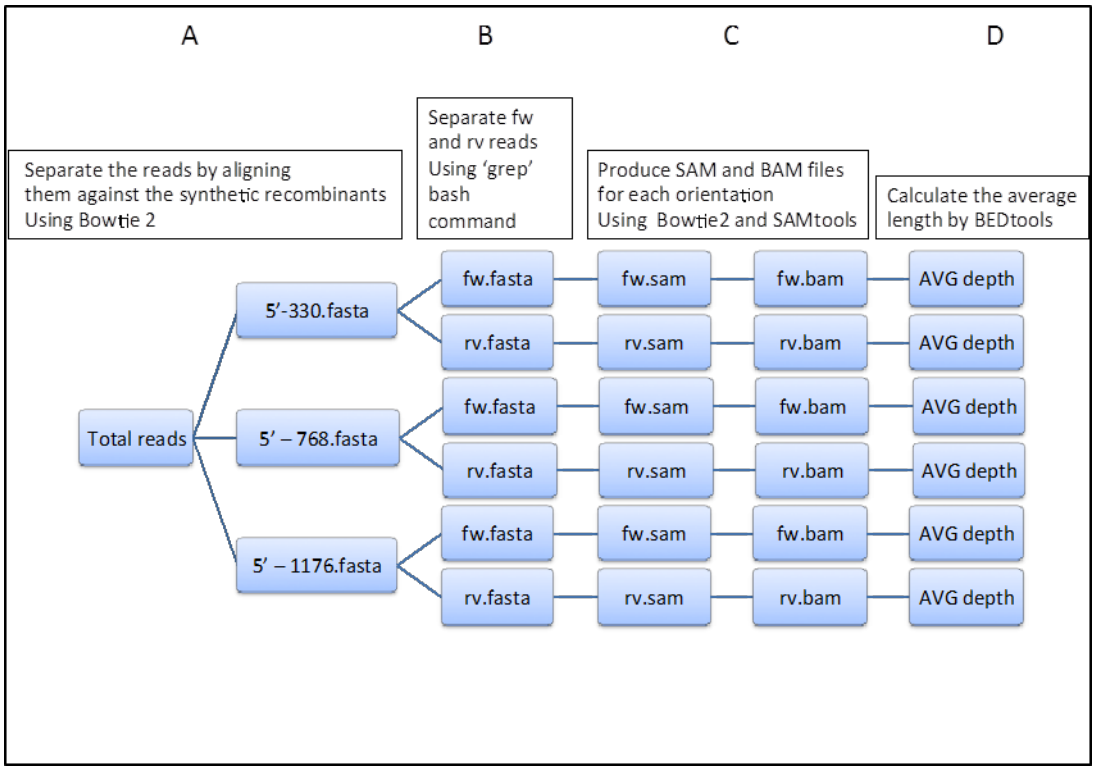


Figure 3-13 Summary of the method used in measuring the depth of coverage
The number of total reads is 4 million when they are treated as single-end A) Reads from the major FASTA file were sorted into their original recombinants sources. B) Forward and reverse reads were separated using the bash command 'grep –A'. C) Re-aligned the forward and reverse reads with the major references (synthetic recombinants) separately to generate SAM files for each of them. The SAM files were then converted into BAM files using 'SAMtools' D) Finally, BEDtools was used to measure the average of depth of coverage from the bam files. fw = Forward reads, rv = Reverse reads. AVG depth = Average of Depth Coverage.

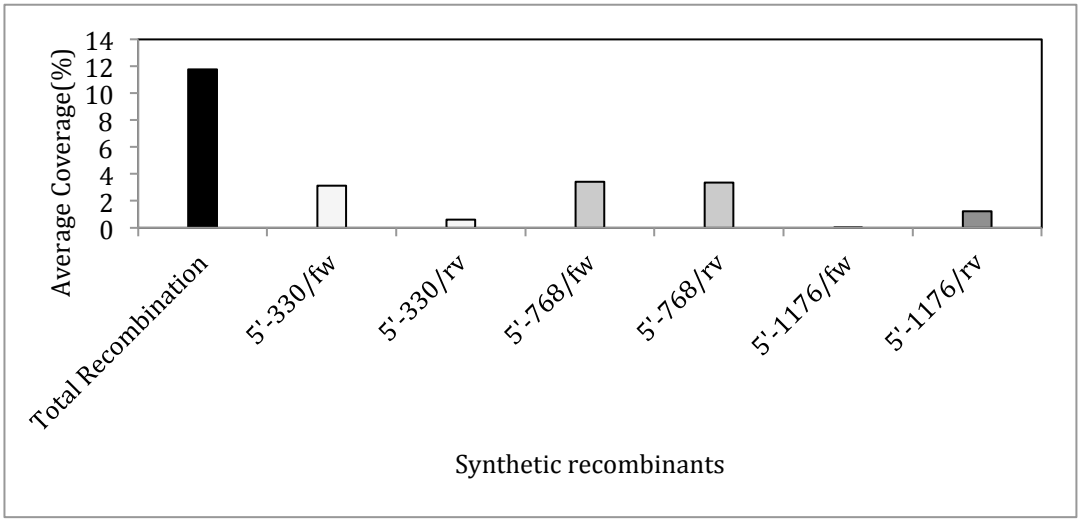


Figure 3-14 The coverage profile of the forward and reverse NGS reads of the synthetic recombinants
Three patterns can be observed based on the distance between the junction location and the ends. The x-axis represents the name of recombinants and the orientation of the reads. The y-axis represents the average of depth of coverage. The black bin refers to all the reads that spanned the junctions within the whole population (3 synthetic recombinants).

3.7.2 Analysing the effect of the junctions' location in the NGS reads on the sensitivity of ViReMa

The position at which the read spans the junction adds yet another challenge for ViReMa in identifying recombination junctions. This stems from the mechanism used by ViReMa to find the junction, *i.e.* extracting 25 nts fragment as a seed (see section 3.5) from the beginning of the read sequence and aligning it to the reference genomes (Langmead et al., 2009). The anchoring of the seed (set to 25 nts in this study) to the reference genome is the signal for ViReMa to start the searching process. Therefore, reads that span the recombination junction within the first 24 nts will not be identified during the analysis. It is 25 undetectable locations, because if the junction locates on nucleotide 25 of the seed it will be detectable by ViReMa. If we applied this on 300 nts read length, the length chosen for this study, a recombination junction occurring at the first or last 24 nts would not be detected leaving 252 possible detectable sites. Consequently, the maximum theoretical sensitivity of recombination detection would be 252/300*100 = 84% (Equation 3-3 A). However, this could be reduced by variables such as biases in the fragmentation of the amplicon with regard to the location of junctions in the amplicon.

To further investigate this, ViReMa was used to analyse Dataset I in Table 3-1. The fragmentation of this dataset was set to precisely 500 nts, so the effect of the DNA shearing could be clearly monitored. The sensitivity of ViReMa for detection each junction was calculated by dividing the number of reported reads on the number of unreported reads that spanned the junction (Figure 3-15), and confirmed with the theoretical sensitivity equation (Equation 3-3). The variability in sensitivity is due to two factors, the distance between the junction and the terminals of the amplicon and the fragmentation length. For example, the junction in recombinant 5'-768 can be calculated by subtracting the seed lengths from the length of the read and dividing the resulting length on the original length of the read (Equation 3-3 A). The forward and reverse reads will have the chance to reach the maximum sensitivity (84%) by both reads equally, as it is located in the middle of the amplicon; away from any possible physical restriction caused by the fragmentation (Figure 3-15).

However, the case is different in recombinant 5'-1176. Since the junction in recombinant 5'-1176 is only 124 nts away from the 3' end of the amplicon, and the fragmentation length is 500 nts, only 124 unique reverse reads can represent the junctions. This because the fragmentation that happened at any of those 124 positions would result in a fragment that contains the junction. Of those 124 unique reads only 124 – (seed length 24) = 100 positions could be detected by ViReMa. The forward reads would never have the chance to reach this junction if all the resulted fragments are exactly 500 nts (which is the case in this analysis) (Figure 3-16 A). However, this can deviate in the real experiment, resulting in slightly shorter or longer fragments (controlled by the standard deviation of the fragmentation method). Theoretically, this would allow a few forward reads to reach the junctions in the shorter fragments. A little modification of Equation 3-3 A is needed to calculate the theoretical sensitivity in for recombinant 5'-1176. As only the reverse reads would span the junction in this case, thus, the seed does not need to be multiplied by 2 anymore. In addition, two new variables are considered, 'r2', which is the maximum distance the reverse reads can reach when they start at the 3' end, and the distance between the junction location and the 3' end "Jd" (Equation 3-3 B). Thus, the theoretical sensitivity of ViReMa to detect the junction of recombinant 5'-1176 is 32% for the reverse reads, and not applicable for the forward reads. However, based on the fact that the read and fragment length can be variable in the real experiment, this equation can be modified to include the variability, which was not investigated in this study. Rather it is based on the assumption that the majority of the fragments and reads would be 500 nts and 300 nts length respectively.

Examining recombinant 5'-330 on the other hand, the forward reads would have different sensitivity from the reverse reads. The physical space in this situation is enough for the forward reads to cover the junction with 84% sensitivity while the reverse reads will be limited by the fragmentation fragment size (500 nts) (Figure 3-16, B). This will allow 130 nts of the reverse read to extend beyond the junction and of these only 130 - (seed length 24) = 106 nucleotides would be detectable by ViReMa. This can be reflected by another modification on Equation 3-3 A by adding the fragmentation factor "f1" (Equation 3-3 C) and the junction location 'J'. Thus, ViReMa theoretical rate sensitivity will be ~35% for the reverse reads (Figure 3-15).

A)

$$TSVt = \frac{r - (S * 2)}{r} * 100$$

B)

$$TSVr = \frac{\left(r - (r2 - Jd)\right) - S}{r} * 100$$

C)

$$TSVf = \frac{\left(r - (f1 - J)\right) - S}{r} * 100$$

Equation 3-3 ViReMa theoretical sensitivity
A) TSVt= Major Theoretical Sensitivity of ViReMa, B) TSVr = Theoretical Sensitivity of ViReMa when the junction is located at the 3' end of the amplicon, r means a factor related to the reverse read was added to the equation C) TSVf= Theoretical Sensitivity of ViReMa, when the junction detectability is partially limited by the fragmentation, f means fragmentation factor was added to equation A to become equation C. S = seed length. J = junction's location. f1 = the largest fragment size possible. r = read length**,** S = seed length. r2 = the limit of the reverse read starts at 3' location. Jd = the distance from the junction location to the 3' end.

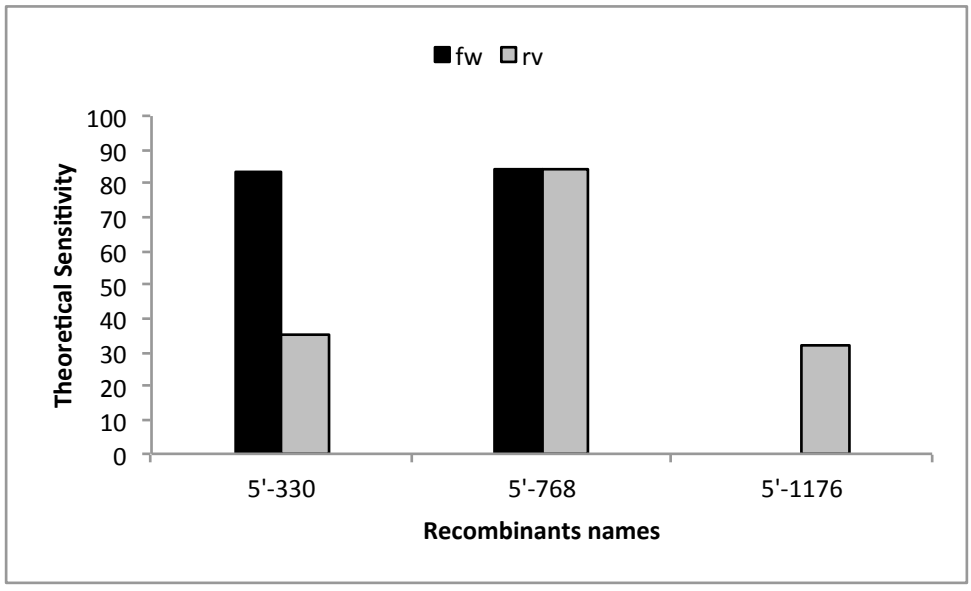
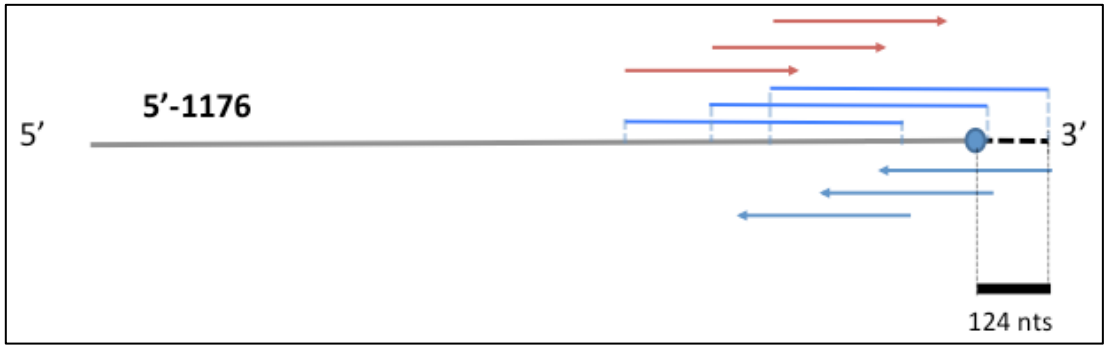


Figure 3-15 Measuring ViReMa theoretical sensitivity of the synthetic recombinants
From the reported reads and by using Equation 3-3 ViReMa theoretical sensitivity was calculated for both read orientations in the three synthetic recombinants (**Figure 3-3**). The y-axis represents the percentage of theoretical sensitivity and the x-axis represents the name of the recombinants. The labels at the top denote the read orientation, fw = forward read, rv = reverse read.

A



B

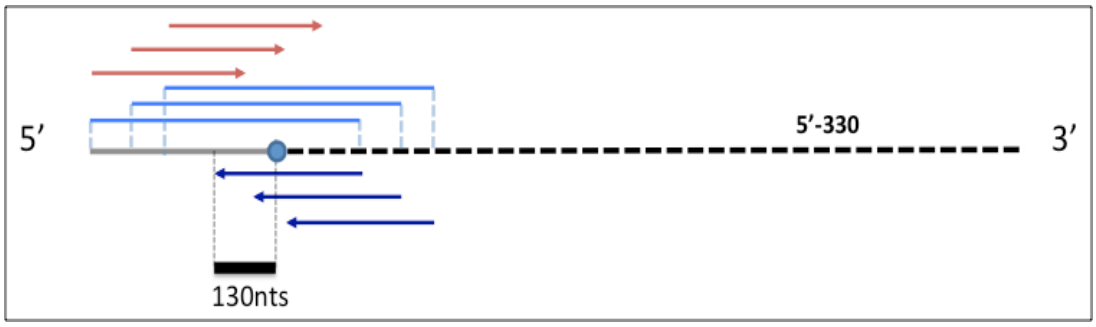


Figure 3-16 Schematic drawing showing the effect of the junction's location in the NGS reads over the ViReMa theoretical sensitivity

The physical space of the reads covering the junction is affected by fragmentation and their closeness to the ends. The bright blue lines represent the possible fragments resulting from fragmentation. To facilitate the presentation these were reduced to three; however, in the real experiment the possibilities can be all the nucleotides before junction. For example there are 124 possible fragmentation sites that can result fragments within which the junction is included in the case of 5’-1176 (upper panel) and 130 sites in the case of 5’-330 (the lower panel). The locations of these possible sites are depicted by the black boxes. The arrows indicate the reads, forward (red) and reverse (dark blue). The circles refer to the junction. The main recombinants are depicted by the long horizontal lines; composed of two parts (grey, dashed-black) reflecting the parent. A) Recombinant 5'-1176. B) Recombinant 5'-330.

3.7.3 Analysing the effect of identical sequences on the sensitivity of ViReMa

The exact site of recombination may be located in a region that has identical nucleotides between the two parents (so-called ambiguous regions). These ambiguous regions happen when nucleotides immediately upstream of the junction in the acceptor virus are identical to the nucleotides immediately upstream of the junction in the donor virus. When junction falls into an ambiguous region, it will be impossible to decide where the actual recombination could have occurred.

Poliovirus type 1 and type 3 are 78% similar within the 1.3kb targeted fragment (from nt 3235 to nt 4548) analysed within this study. To gain a general insight into the ambiguous regions shared by both viruses, Jdotter (Brodie *et al.*, 2004) was used to plot the similarities of the targeted region between the two viruses in a dotplot presentation using the default setting of sliding widow, 8 nts (Figure 3-17). The plot showed some ambiguous regions between the genomes, appearing as a continuous diagonal line when the two viruses share the same sequence and discontinuous over the divergent regions. From the dotplot it could be observed that the diagonal line mostly continuous in the *2A*- and *2C*-encoding region, which reflects the presence of long similar stretches over these regions. The junction location would not be identifiable if it lies within these ambiguous regions. Instead, ViReMa would push the potential junction towards the 3' end of that ambiguous region (this can be set to 5' of the ambiguous region). By doing this, ViReMa may report several reads, which could have originally derived from different junctions, for one recombination junction (Figure 3-18). Note: the expressions 'ambiguous regions' and 'identical sequence' are used interchangeably hereafter as they refer to the same concept.

Analysing the presence of identical sequences in more detail, it was found that these regions are distributed as short regions of identities between 2 and 21 nts in length in the P2 region and they exist in different frequencies. For example, there are 99 identical sequences with 2 nts length (dinucleotide) scattered over the targeted region while only one identical sequence with 21 nts length lives in the *2C*-coding region. In contrast, there are ~284 divergent nucleotides between poliovirus type 1 and type 3 within the same region.

To investigate how these short identical sequences would affect ViReMa detection power, simulated libraries – composed of precise and imprecise recombinants at every possible location of the amplicon - were used. Precise recombinants are those recombinants that occurred at a comparable site between the parental genomes *i.e.* the resulted recombinant retain the same genome length of the parental genomes. On the other hand, the imprecise recombinants are those that occur at unrelated site, as a result; the recombinant virus will contain sequence duplication or deletion (see sections 1.3.5)

The libraries were simulated by a custom-written Perl script (Appendix 1) that uses the junctions' locations from both parents to generate the recombination read without any mutation (Table 3-1 Datasets J1, K1), or with 1% random mutation (Table 3-1, J2, K2) to study the effect of mutation in the context of the short identical sequences. The libraries contained 1252 unique junction reads as the 24 undetectable nucleotides from the terminals were removed. ViReMa was then used to find the recombination junctions in all libraries using N0X5 mode.

In the case of precise recombinants, the results showed that ViReMa could detect ~23% junctions from the 1252 total junctions when no mutations were applied on the library. The percentage of detectable junctions were reduced from ~23% to ~20% when 1% mutation was applied to the library (Figure 3-19) Analysing the results, it was found that ViReMa found the recombination junction in all the reads. However, because it was not possible to allocate the exact junction location if it occurred within an identical stretch, ViReMa reported several unique junctions as a junction at a single location. For example, ViReMa reported 22 unique junctions within a 21mers identical sequence as one unique junction. Interestingly, the reported junctions constituted 23% of the total junctions, which is the same divergence ratio between the two viruses. This reflects the critical role of these divergent locations in helping the algorithm detect the nearest location to the exact junction location.

In the case of imprecise recombinants, ViReMa could detect more recombination junctions than in the precise recombinants case, ~67% junctions from the 1252 total junctions when no mutations were applied on the library and ~44% when 1% mutation was applied to the library (Figure 3-19). As these recombinants are imprecise, they occurred at unrelated sites, thereby decreasing the effect of the

identical sequence, which in its turn would increase the detectability of ViReMa. This is discussed further in Chapter 5 and the correlation between the numbers of recombination unique reads and the length of the identical sequence is demonstrated in detail in Figure 5-18.

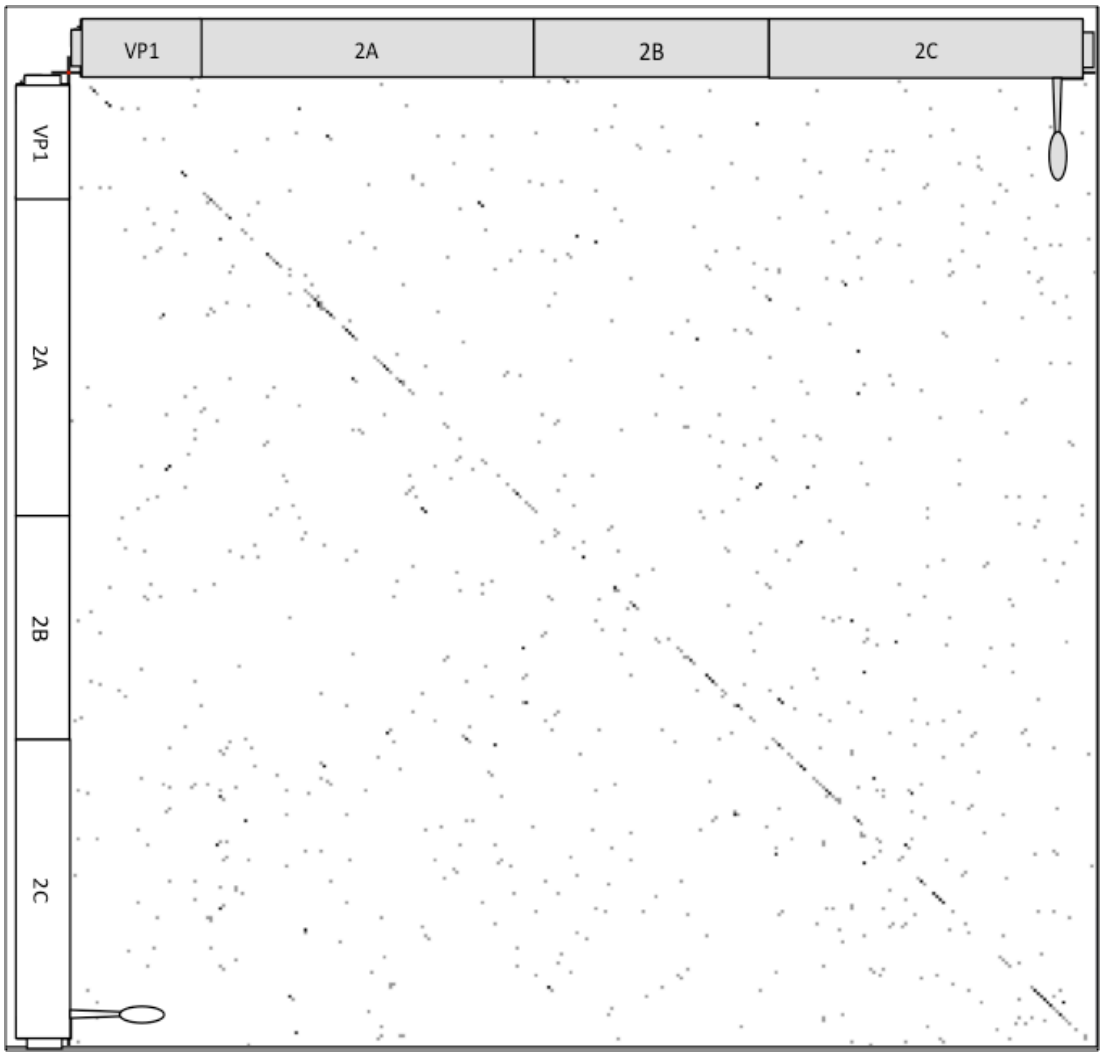


Figure 3-17 Dotplot shows the identical sequences between the two viruses graphically
A dotplot presentation displaying the identities between PV1 and PV3 within the targeted region; the PV3 sequence runs along the x-axis and the PV1 along the y-axis. They are demonstrated as drawing genomes within which the protein encoding regions are indicated (the white colour refers to the PV1 genome and the grey to the PV3 genome). In regions where the two viral sequences are similar to each other, a row of high scores will run diagonally across the dot matrix.

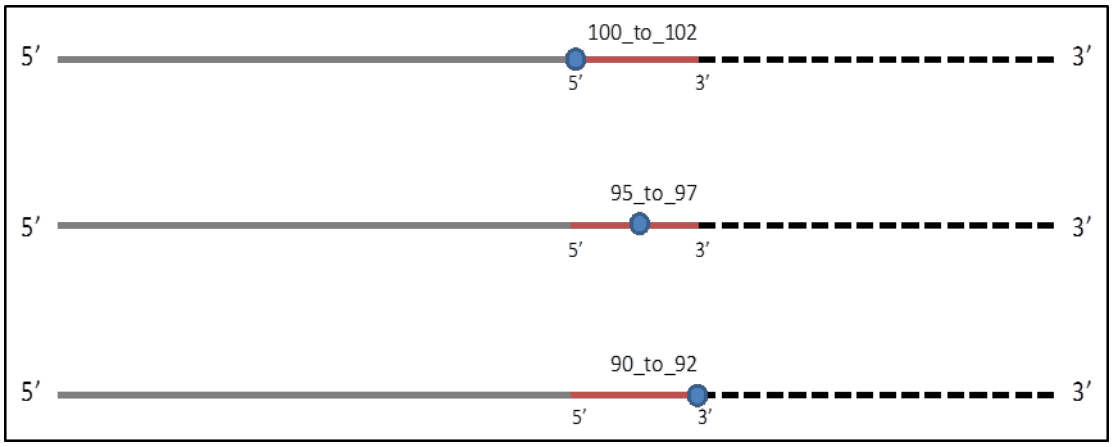


Figure 3-18 schematic illustration of reporting junction location within an identical sequence
When ViReMa encounters a junction located in an ambiguous region, it pushes the junction location towards the 3' end of the ambiguous regions. The long horizontal lines represent three different recombinants which have their junctions located within an identical sequence; the grey and dashed-black lines indicate the parental genomes of each recombinant. The red thick lines denote 10 nts ambiguous regions (*i.e.* a stretch of identical nucleotides between the two viruses), where the circles represent the actual junctions. The numbers above the ambiguous regions refer to the location at which the polymerase switched from the first virus (black-dashed line) to the other virus (grey line). If the three recombinants occurred in a real co-infection reaction, ViReMa would not be able to differentiate between the three types of junctions; thus, will report them all as one junction, using the 3' end of the ambiguous region as a reference location. In the case above ViReMa would report the three junctions with the location 90_to_92.

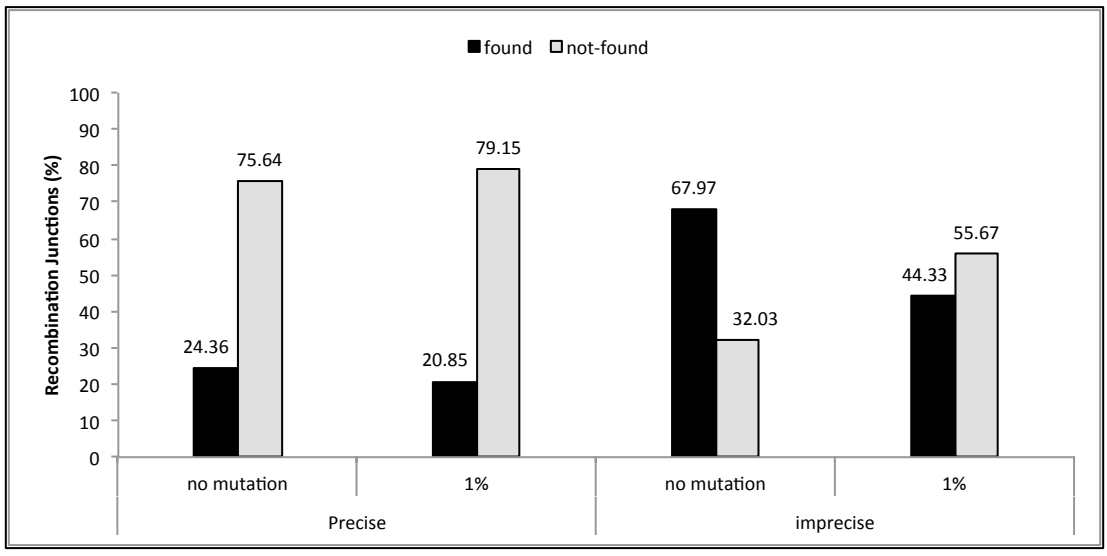


Figure 3-19 Identical sequence effect on ViReMa (precise and imprecise)
The sensitivity of ViReMa in detecting unique junction across the whole targeted genome within different NGS datasets. The x-axis shows the total recombination junctions and the detected ones while the y-axis refers to the percentage of recombination events. The colours refer to the results of ViReMa, either detectable reads (black), or undetectable (grey).

3.8 Mathematical function to calculate the theoretical sensitivity

Depending on the ViReMa theoretical sensitivity, a function that can predict the sensitivity over the amplicon was established. The function calculates, for a particular junction at a particular location, the percentage of how many detectable junction sites in the NGS reads are available for ViReMa to detect, taking into consideration the factors that affect the physical space, such as fragmentation and distance from the ends of the amplicon.

The function (Equation 3-4) was established based on four different locations that are important in defining the physical space of the reads to reach the junctions. The first location (termed r1) equals 300 which is defined by the maximum location the forward read can reach, the second location (termed f1) equals 500, and is defined by the first 500 fragments generated by the fragmentation process with respect to the 5' end of the amplicon. The remaining two locations, which were termed f2 and r2, follow the same concept from the other side of the amplicon (Figure 3-20- black circles). Because the theoretical sensitivity of the reverse reads is a mirror profile of the forward reads, they can both be calculated from the same functions by normalising the former to the latter. This can be done by subtracting the junction location (J) from the length of the amplicon (Amp - J = normalised reverse reads). For example, if the junction is located at nucleotide 1000 on the amplicon, then the criteria of the third component in the function is satisfied, thus it can be used to calculate the theoretical sensitivity of the forward reads. As for the reverse reads, this needs to be normalised to become 1300 – 1000 = 300, which will fulfil the criteria of the second component of the function.

With the assumption that the coverage of reads is sufficiently high, this function can be used to determine the location on the amplicon at which ViReMa can start and stop detecting. In addition to that, it can predict the theoretical sensitivity at any location, which can be used as a reference profile to interpret the distribution of recombination reads from the real experiment.

The results in Figure 3-20 demonstrate the ViReMa theoretical sensitivity calculated by the function (Equation 3-4) with an amplicon size of 1300, read length of 300 nts, fragmentation length of 500 nts, and a seed length of 25 nts. It can be observed that

ViReMa will start detecting more than one unique read for the forward reads at nucleotide 26 (0.3%), and will increase to reach the highest achievable sensitivity ~84% at nucleotide 274. This happens because the junction at this location can be spanned by all the nucleotides of the reads, and ViReMa can detect it in all the nucleotides except in the last and first 25 nucleotides. The sensitivity will remain constant till the junction reaches nucleotide 826, where the effect of fragmentation starts to appear. From this point onwards the sensitivity will decrease until it reaches the 0.3% again at nucleotide 1074. Based on the read and fragment length, the forward reads cannot cross any junction located after this position (Figure 3-20, red line). The same scenario applies to the reverse reads, but from the other side of the amplicon (Figure 3-20, blue line). The first and last 25 nts of the amplicon will remain undetectable.

$$f(J) = \begin{cases} \dfrac{r-(r1-Jd)-S}{r} * 100 & if\ S \le J \le r1 - S \\ \dfrac{r-(2s)}{r} * 100 & if\ \ r1 - S \le J \le f2 \\ \dfrac{r-(J-f2)-S}{r} * 100 & if\ f2 + S \le J \le f2 + (r - S) \end{cases}$$

Equation 3-4 Mathematical function to calculate ViReMa theoretical sensitivity
The function is composed of three components; the first calculates the sensitivity if the junction lies in the first 300 nts of the amplicon. The second component specialises in calculating the sensitivity of the junctions when they are located in the middle of the amplicon, and the third calculates the sensitivity for junctions located at the last part of the amplicon. The function is plotted in (Figure 3-20 - red line). The forward reads can't span any junction after nucleotides 1075 because of the physical restrictions imposed by the fragmentation and the closeness to the 3' end. r = the read length. r1 = The furthest location the forward read can reach is when it starts from the 5' end of the amplicon. J = the junction location. S = The seed length. f2 = the furthest location from the 3' end at which the fragmentation process could cut. r and r1 are both 300, but they are referred to as such to differentiate between their technical implications, r1 is a distance, r is a length. Jd = the distance from the junction location to the 3' end.

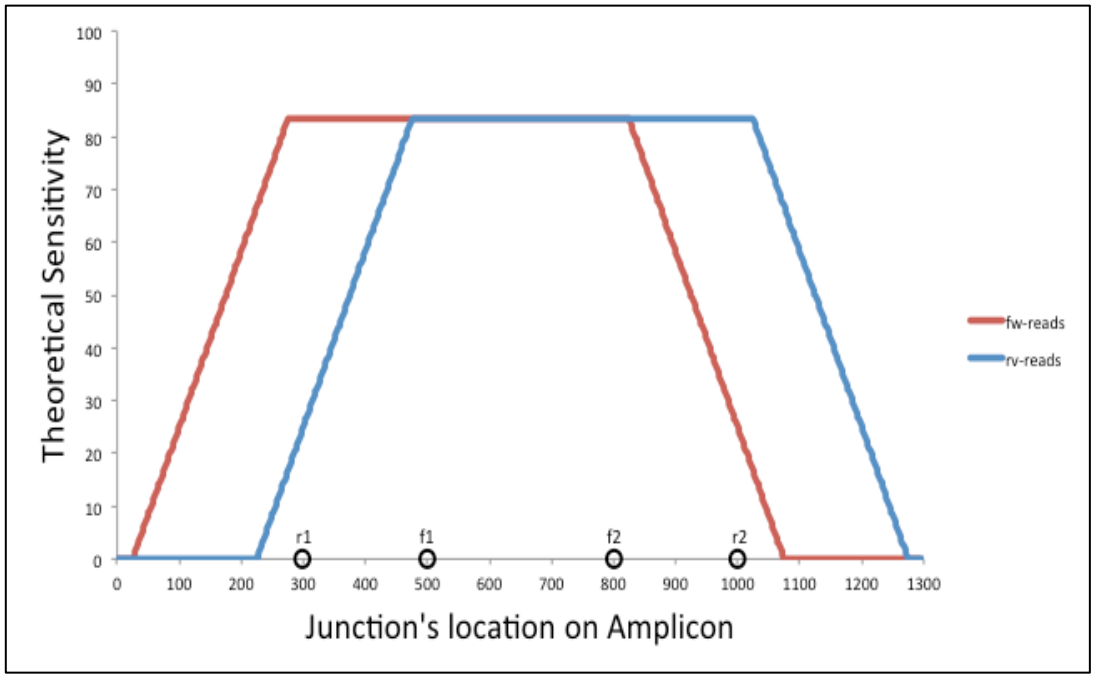


Figure 3-20 Diagram shows ViReMa theoretical sensitivity calculated by the function for the whole amplicon
The function (Equation 3-4) created two patterns of sensitivity; one for the forward reads (red line), and the other for the reverse reads (blue line). The black circles indicate the major locations the function used to generate the sensitivity. X- axis represents the location over the amplicon and the y-axis refers to the theoretical sensitivity (%)

3.9 Discussion

A bioinformatics analysis system was developed to identify the recombinants from a co-infection between two serotypes of poliovirus. The major hurdle in analysing the NGS datasets that derived from viral recombination pool lies in mapping the reads that spanned the recombination junctions of either parental genome. To overcome this, the Python application ViReMa (Routh and Johnson, 2014) was used to identify recombination junctions within different simulated NGS datasets.

NGS factors were optimised through data simulation – which would be considered later for the sequencing of the real samples - to guarantee high-resolution output that allows ViReMa to perform at its best. This includes the number and length of NGS reads, the type of NGS sequencing used, and the types of reads whether paired-end or single-end. The conditions were decided upon based on the aim and resources of this project, and more importantly on the knowledge of recombination occurrence in polioviruses. For instance, the region of interest was chosen based on a previous study from our lab using an artificial system (termed the CRE-REP assay) to isolate recombinants between poliovirus type 1 and type 3 (Lowry et al., 2014). Moreover, this region is characterised by being 23% divergent between the two viruses, which would facilitate the identification of the actual recombination junctions.

However, the judgement of the location at which the recombination events occurred is highly correlated with the amount and type of mutations that could present in the final datasets. Therefore different types and rates of mutations (Figure 3-5) were applied onto the datasets (Table 3-1) and ViReMa with different settings (modes - Table 3-2) used against each in order to identify the recombination reads. Subsequently, the sensitivity and specificity of ViReMa performance was measured by the ROC curve (Figure 3-10). These results in the ROC curve were produced from three different simulated datasets with different rates of mutations. Nonetheless, the analysis in later stages was focused on the datasets that contained 1% mutation, as this is the closet one in nature to the dataset produced by the real experiment. The other datasets were used for the sake of comparison to improve the analysis pipeline.

When analysing the results of dataset C (1% mutation) from the ROC curve, it was found that the best performance for ViReMa was achieved through the mode N0X5

(~64%). This mode does not allow any mismatch to happen either in the seed or in the remaining part of the read. Additionally, it does not depend on the length of similarities around the junction ('X parameter'). Even though the 'X' is set to 5 in this mode, it will be disallowed by adjusting the 'N parameter' to 0. On the contrary, mode N1X5, which was found to perform slightly better regarding the sensitivity, does depend on the similarities around the junction. The reason of this increase in sensitivity can be explained by the fact that N1X5 mode allows 1 mismatch when aligning the reads. This would lessen the chance of rejecting the recombination read based on the presence of mismatch. However, this would happen at the expense of specificity; as mode N1X5 depends on the length of similarities around the junction, it would report some false positive junctions. Particularly when the algorithm encounters a far longer identical stretch than 5 nts, as illustrated in Figure 3-9. Because one of the main goals of this study is to understand the molecular determinants of recombination, the specificity was favoured and accordingly, mode N0X5 was chosen for the rest of the analysis.

Using ViReMa with mode N0X5 to search recombination reads in a dataset with 1% mutation would be able to detect ~64% of the recombination reads (those spanning the recombination junction) and ~36% would be undetectable. These 36% reads were found to be either reads with mutation within 25 nucleotides of the junction or reads that spanned the junction at their first or last 25 nucleotides. In addition to this, a small proportion (~0.1%) of reads containing incorrect junctions were reported by ViReMa (Type 3, section 3.6.1 and read D, Figure 3-11). These ~0.1% reads were scrutinised and found to have a mutation right next to the identical sequence where the exact junction had occurred. Consequently, this would cause ViReMa to skip the exact junction and report a junction that was separated by an identical sequence from the exact one. The ability to detect these reads in the simulated dataset stems from that fact that the number of simulated recombinants is limited (three recombinants), thus any reported junction apart from the original three is recognisable. However, this is not the case in the experimental data as the identity and the number of recombinants are not known.

The mutation that caused the third type of erroneous reads can be inherited from one of three possible sources throughout the experiment, either during the virus

replication in the cell, during the PCR amplification, or during the sequencing. Although with the presence of three sources of mutation, the probability of the occurrence of such a mutation is very low based on the mutation rate $5.2*10^{-4}$, and therefore it can be speculated that this sort of error would not have a drastic effect on the specificity. It can be argued that the mutation may happen at an early cycle during the PCR, which might give it a chance to be amplified in the dataset. Nonetheless, these mutations should happen near to an identical sequence to be able to mislead the ViReMa, and the vast majority of identical sequences between the two viruses are of 2 nucleotides long. This means that the shift caused by this mutation between the correct and incorrect junctions would not be greater than 3 nucleotides for the most part, which can still be considered very close to the correct junction site. Therefore, this would not change the sequence composition of the surrounding area of the junction, which would be studied as an influential factor for recombination in chapter 5.

Having determined the specificity and sensitivity of ViReMa in finding recombination reads, the sensitivity of ViReMa in finding the recombination junctions across the whole amplicon was tested. That is, if the recombination junctions could occur at every nucleotide in the 1.3kb amplicon, how many of those could the ViReMa detect? Using a library containing unique reads of all the possible precise and imprecise junctions ViReMa was able to detect 23% of all the precise junctions in comparison to ~67% of the imprecise junctions. Although it recognised all reads in these dataset as potential recombinants, ViReMa could not report the exact location for all the reads as most of them lie within sequences identical in the aligned parental genomes. Accordingly, it pushes the location towards the 3' end of the identical sequence as illustrated in Figure 3-18, which can generate a bias towards the long identical stretch. The ability of ViReMa to detect more imprecise recombinants reads can be explained by the fact that these molecules are imprecise. That is, the 5' and 3' side of the junction would include sequences from unrelated regions in the genomes, thus creating more distinguishable sites. In the real experiment, the number of reads of each junction can be considered a reflection of the replicative nature of the recombinants. However, this should be analysed carefully to differentiate between high representations of reads caused by viable fit recombinants, and those caused by systematic bias (see chapter 5).

On the other hand, the location of the junction in the amplicon and the NGS read was found to be a major player in determining the sensitivity. It was found that the closer the junction to the amplicon's termini the less coverage it received (Figure 3-14). Based on this, the theoretical sensitivity of ViReMa would decrease as the physical space available for the reads to span the junction is limited at the termini (Figure 3-16). This can be used in the data produced by the real experiment to differentiate between rare recombinants occurring in regions unfavourable for recombination from those likely underrepresented in the data analysis.

ViReMa was found to perform efficiently in identifying recombination junctions amongst NGS simulated datasets. Although the bioinformatics analysis demonstrated here is efficient in detecting recombination junctions in the P2 region of poliovirus, with a few modifications it can be accommodated to study recombination in the rest of the poliovirus genome and other viruses.

4 Poliovirus *in vitro* intertypic recombination

4.1 Introduction

Many recombinants have been isolated and characterised from field samples following outbreaks of infection (Dedepsidis et al., 2010, Zhang et al., 2010, Kew et al., 2002). Previous studies suggested potential factors that could influence recombination such as RNA secondary structure and specific sequence motifs (section 1.3.3). Nevertheless, the mechanism by which recombination occurs is still poorly understood.

Recombination is a rare biological event as it occurs at a low frequency within the virus population. The frequency of intertypic recombination between poliovirus type1 and type3 recombinants was found to from $2*10^{-4} – 1*10^{-5}$ recombinants molecules from the total virus progeny (Tolskaya et al., 1983). In contrast, intratypic recombination was found to occur 10-100-fold more frequently (Kirkegaard and Baltimore, 1986, Lowry et al., 2014). This rarity of recombination coupled with the inevitable competition of recombinants with parental viruses has been the major difficulty in isolating recombinants from a population after mixed infection.

Early *in vitro* studies have largely relied on using viruses possessing different selectable markers. The main approach used in these methods is to carry out the co-infection under permissive conditions with two different parental viruses bearing genetic markers such as drug resistance or temperature sensitivity. Those that escape the selection process would be recombinants virus (Kirkegaard and Baltimore, 1986). Although these methods have largely contributed to the understanding of recombination, their success is conditioned by the production of recombinants at a higher frequencies than the reversion rate which was found to be 10^{-3}-10^{-4} (Holland et al., 1982). Recently, several methods have been proposed to study recombination in a more efficient way. Nonetheless, the vast majority of this study focused on intratypic recombination. For example Jarvis & Kirkegaard suggested a system to study intratypic recombinants in the absence of genetic selection markers (Jarvis and Kirkegaard, 1992). The system they proposed uses PCR-based technique not dependent on the viability of recombinants. Furthermore, Runckel *et al* developed a more inclusive experimental system to study intratypic recombination with a higher

resolution map, i.e., isolate recombinants from all the genome rather than from a particular region. In their method, they have utilised a synthetic poliovirus genome (type 1) and next generation sequencing to obtain a wide range of recombinants. However, due to the reliance on serial passage, specific PCR, and amplicon size selection, the system could detect precise viable recombinants only (Runckel *et al.*, 2013).

In an attempt to study intertypic recombination, Lowry *et al.* have developed a novel reverse genetic approach with the parental genomes being incapable of producing any viable progeny but recombinants (Lowry et al., 2014). Upon co-transfection into permissive cell line the recombination should occur within the region between the beginning of VP1 and the CRE to produce a viable recombinant virus. The absence of CD155 receptors on the permissive cell would prevent the recovered virus from re-infecting new cells. This can be translated into two important features. Firstly, it ensures that recombinant viruses are isolated and characterised as near to the recombination event as possible. Secondly, as the produced viruses cannot go back to the cells; therefore no competition is allowed apart from the one that took place inside the cells during the replication. By minimising the selection, the CRE-REP allows the isolation of a range of possible recombinant viruses could be recovered regardless of their growth advantages.

The CRE-REP system demonstrated that recombination is a biphasic process in which an initial recombinant is followed by a resolution process, deleting insertion sequences and optimising virus fitness. The initial recombinants were found to contain insertions which were described as 'imprecise' recombinants whereas the resolved products were described as 'precise' recombinants. Despite the fact that only imprecise recombinants with insertions were detectable in the CRE-REP assay, recombinants could also theoretically contain deletion as would be discussed later in this chapter and the following ones. In this study the same names suggested by Lowry *et al*. were used with little modifications (see section 2.5); the recombinants with insertion were called 'imprecise-insertion', and 'imprecise-deletion' if they contained a deletion. If no insertion or deletion – relative to the parental genomes - were involved they were termed 'precise' recombinants (Figure 4-1). Since the CRE-REP assay involved harvesting virus progeny 24-48 hours post infection and

isolating by limit dilution in HeLa cells, it was hypothesized that isolating recombinants at an even earlier stage would decrease the selection and allow detection of a wider range of recombinants. Therefore, a viability-independent assay involving PCR and NGS sequencing was developed in this study to amplify intertypic recombinants between wild-type poliovirus type 1 and type 3 in the absence of selection pressure. For the sake of comparison the targeted region in this study was the same that was targeted by the CRE-REP assay.

Aim

Based on the findings from the CRE-REP assay, the main aim was to develop a more natural and inclusive PCR-based screening strategy that a) uses co-infection strategy rather than co-transfection, which would allow for the natural virus infection to take place, rather than bypassing this by introducing a naked virus RNA into cells b) uses wild-type viruses rather than mutants contain non-virus sequence d) can detect recombinants regardless of their viability. This was planned with the intention of then using this assay as the basis for an NGS study

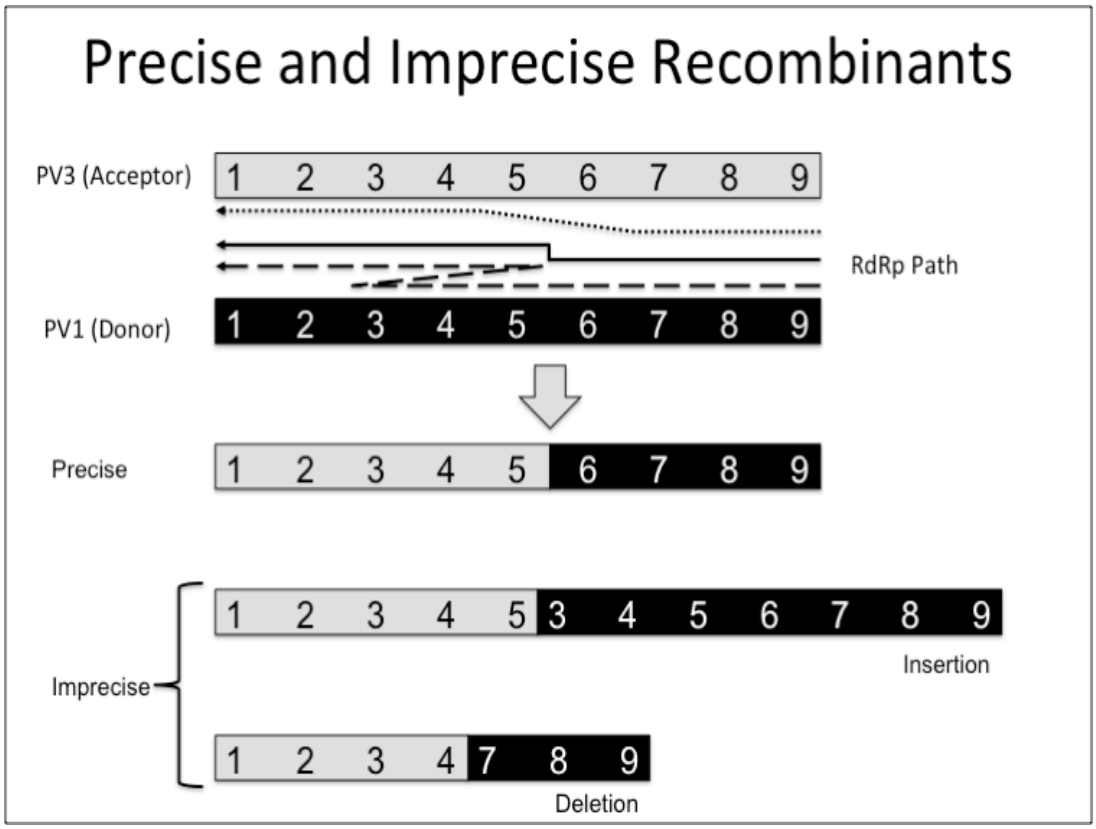


Figure 4-1 Schematic drawing illustrates the types of recombinants (precise, imprecise-deletion, imprecise-insertion)
Intertypic recombination between poliovirus type 1 and type 3 produces two types of recombinants, precise, and imprecise. The latter is further divided into imprecise-deletion and imprecise-insertion. The rectangles represent the viral genomes while the colours denote the type of the virus. The arrows between the parental genomes rectangles represent the RdRp path, if the RdRp switched template as the dotted-line path then imprecise-deletion would be generated, the solid-line path would generate precise recombinant, and the dashed-line path would result in an imprecise-insertion recombinant.

4.2 Proposed experimental system

A replication-dependent experimental method was developed to isolate and analyse recombinants 5 hours post infection to minimise the effect of selection pressure. The proposed system was based on three essential elements:

A) Increasing the virus progeny to enhance the opportunities for recombination detection. As recombinants exist at a low frequency; increasing the scale of the infection increases the absolute number of recombinants in the sample, hence the recombinants are more likely e to be detectable.
B) Optimising the time to harvest the virus progeny in order to detect a wide range of recombinants.

C) Efficient extraction of the viral RNA in order to enhance recombinants' representation in the samples by preventing any possible loss happening in the extraction method.

One of the biggest challenges was the low frequency at which intertypic recombination exists. Increasing the yield of the co-infection experiment was vital in bringing the absolute number of recombinants into the detectable range of the PCR assay used in this study. To achieve this, four T175 flasks were co-infected in parallel at a high multiplicity of infection (10 MOI), the cells were then pooled together ($7.6*10^{7}$ cells in total), which was followed by RNA extraction, cDNA synthesis, and PCR amplification.

The next step was to optimise the harvest time of virus progeny to ensure the detection of the recombinants that were present. Because recombinants can be formed as early as 2.5h post infection (Egger and Bienz, 2002) and the virus is released from the cells 6h post infection (Kew et al., 2005) it was reasoned that 5h post infection would be the peak moment to capture recombinants while they were inside the cells. At this stage, the recombinant viruses had a maximum opportunity to be formed in the cell before they were released to the media. Additionally, after 5 hours the virus did not have time to go through another round of infection. Therefore the selection pressure was limited to the ability of some recombinants to replicate while they were inside the cells. This situation maximised the chance to analyse a wide range of recombinants that were present at the time of infection - these can be categorised into five different types.

1. Encapsidated precise recombinants. Some precise recombinants maintaining protein-protein compatibility that were capable of being both translated and replicated.
2. Free RNA of precise recombinants. These recombinant molecules - with no deletions or insertions - either did not have time to be encapsidated or are of a size or junction type that was incompatible with encapsidation.
3. In-frame imprecise recombinants. These recombinants maintain the open reading frame and could carry either deletions or insertions. The recombinants with the former would be shorter and most probably would not be able to replicate any further as they had lost part of their genome (see

section 6.2.3) while the genomes of the latter might be able to replicate and resolve into precise recombinants (Lowry et al., 2014).

4. Out-of-frame imprecise recombinants. These recombinants did not maintain the open reading frame and could carry either deletions or insertions. In both cases these recombinants could not be translated or replicated.

Finally, in order to extract the viral RNA without any critical loss that would undermine the presence of any of the above-mentioned molecules, an efficient concentrating step was required. This was achieved by collecting the pooled cells in a minimal amount of media, followed by three freeze-thaw cycles to rupture the cells. Subsequently, the cells' debris was removed, and the supernatant, which contained the virus, was filtered. The pooling of virus progeny, from the four large flasks, into a minimal amount of media, concentrated the amount of virus in the sample. At this time during the infection, the viruses would still be inside the cells, and therefore they could be concentrated in the sample by concentrating the cells. That is, the cells served as cushions to save the virus from being diluted by the media.

A schematic drawing showing the experimental steps is demonstrated in Figure 4-2. Briefly, poliovirus type 1 and type 3 were used to co-infect four T175 flasks at MOI of 10 (to ensure that the vast majority of the cells were dually infected (see section 6.2.1)). The cells were incubated for 30 minutes at 37ºC before the virus was removed and the cells were washed with PBS. Fresh media was added and the flasks were incubated at 37ºC for 5h. Subsequently, cells were pooled in 2ml of the same media to be ruptured by freeze-thaw, the cells' debris was removed and the supernatant containing the viral RNA was filtered through 0.2µm filters. Finally, from the filtrate, the RNA was extracted, followed by cDNA synthesis, PCR amplification, and sequencing.

4.3 Optimisation of the PCR reaction

The next step was to optimise the PCR conditions in order to improve the detection of recombinants. Before reaching the final optimised PCR protocol, which demonstrated a high sensitivity in amplifying a low number of cDNA copies (see

section 4.3.2), several elements were examined (Table 4-1). The final PCR protocol used throughout this study is described in section 2.2.11. Briefly, Promega PCR Master Mix containing *Taq* DNA polymerase was used in 50µl reactions and the PCR thermocycler was set to 40 cycles. Unless otherwise mentioned, the quantity of template used was ~0.1-0.3ng in 5µl volume of distilled water, which equalled ~10^7 copies of either the plasmids containing the full-length of the viruses or the cDNA. The copy number was calculated using an online calculator developed by the Genomic & Sequencing Centre of the University of Rhode Island (http://cels.uri.edu/gsc/cndna.html).

Optimising the PCR assay was a crucial step for enhancing detection; optimisation included both specificity and sensitivity. The former – in the context of this study - is the ability of the primers under certain PCR conditions to specifically amplify recombinants from wild-type poliovirus co-infection while the latter is the lowest number of cDNA copies the system can detect. Wide ranges of conditions were investigated during the optimisation of these PCR reactions. These included different primer pairs, different annealing temperatures, and different PCR reactants concentrations such as primers, templates and Mg++ (Table 4-1). Additionally, emulsion PCR was tested as a potential method to increase the sensitivity and specify of recombinants amplification. However, the attempt failed, as this was tried earlier in the project on a smaller-scale infection. For the sake of abbreviation and clarity, the final optimised conditions were discussed in this chapter.

4.1.1 Designing specific primers to amplify recombinants

One of the crucial steps for the experiment was to design primer-pairs specific for amplifying the recombinants that resulted from the co-infection experiment without any significant cross-reaction with the parental genomes. Since this non-specific reaction would produce artefactual products that would look like products from recombinants, the presence of these in the PCR product would confound subsequent analysis. The region of interest within which recombination was targeted to occur was located between the end of the capsid coding region (nucleotide 3235 in the poliovirus genome) and the CRE (nucleotide 4548) (Figure 4-3).

Initially, the population of recombinants that occurred within the P2 region and characterised by involving the dissociation of RdRp from PV1-Mahoney (Donor) and re-association to PV3-Leon (Recipient) was investigated; hence it was termed 'recP2-R_LD_M' (Table 2-3). Hereafter, this abbreviation will be used to refer to this recombinants' population and its corresponding primers pair will be referred to as 'primP2-R_LD_M' (Table 2-3). The analysis of this population would verify the reliability of the current system, which is comparable to what was analysed in the CRE-REP assay (Lowry et al., 2014). It was reasoned that if precise and imprecise recombinants were successfully detected and analysed from this population then the experiment would be considered successful. As a consequence, it could then be expanded to encompass different regions and populations. (Note: the words 'acceptor' and 'recipient' were used interchangeably in this report to refer to the genome on which the RdRp would re-associate.)

The primer pair primP2-R_LD_M (PV3-3235F, PV1-4548R, Figure 4-3), which generated a product of 1313bp, was carefully designed. To assure the sensitivity and enhance the efficiency particular care was given to the 3' end of the primers as they were considered specificity determinant regions (Miura et al., 2005, Onodera, 2007). As the two viruses were ~78% similar (within the targeted region) the primers were designed with their 3' end to include as much non-identical nucleotide as possible, thereby assuring variant anchoring sites to prevent mispriming. Additionally, priority was given to allow nucleotide C' or 'G' to be encompassed within the 3' end to promote binding. Finally, to avoid the formation of primer-dimers the 3' ends of both of the primers were not complementary. The properties of the final designed primers including the melting temperature, the percentage of GC content, and the PCR suitability were tested by The PCR Primer Stats software, part of the Sequence Manipulation Suite (Stothard, 2000). Several primer pairs were designed and the ones that passed all the tests were used in this experiment (Table 2-3).

Condition optimised	Purpose	Optimisation
Templates' concentration	A) This was considered to increase the PCR products. Increasing the number of the DNA template copies would create more chance for the recombinant molecules to be amplified. B) On the other hand, decreasing the template would allow for optimisation of the sensitivity (see section 4.3.2).	In the case of A, this was achieved by increasing the concentration of the DNA. The optimal situation was in the appearance of a high-intensity band with the absence of non-specific product. As for B, this was carried out by using a serial dilution to determine the sensitivity of the detection.
Primers' concentration.	This was adjusted in order to minimise the primer-dimer and the non-specific reactions.	This was achieved by applying different concentrations of the primers. The optimal situation was the highest primer concentration with no appearance of the primer-dimer.
Annealing temperature	This was optimised to increase the specificity.	This was achieved by using a temperature gradient programme in the PCR machine. The best annealing temperature was the one at which the PCR reaction produced the highest intensity specific band with no non-specific products (Figure 4-4).
Elongation time	This was adjusted to be as inclusive as possible i.e. imprecise recombinants with insertion could have a chance of being amplified.	This was achieved by increasing the elongation time without affecting either the specificity or sensitivity.
MG^{2+}	This was adjusted to optimise the specificity. MG^{2+} can enhance the activity of the *Taq* polymerase, which can reflect on the specificity.	Serial dilution of MG^{2+} was used. The optimal situation was for the specific band to appear with no extra non-specific products.

PCR cycle number	This was optimised to allow enough PCR cycles to increase the number of recombinants without producing non-specific products.	This was achieved by trying several cycle numbers between 30-40.

Table 4-1 Description of the conditions optimised for the PCR reaction

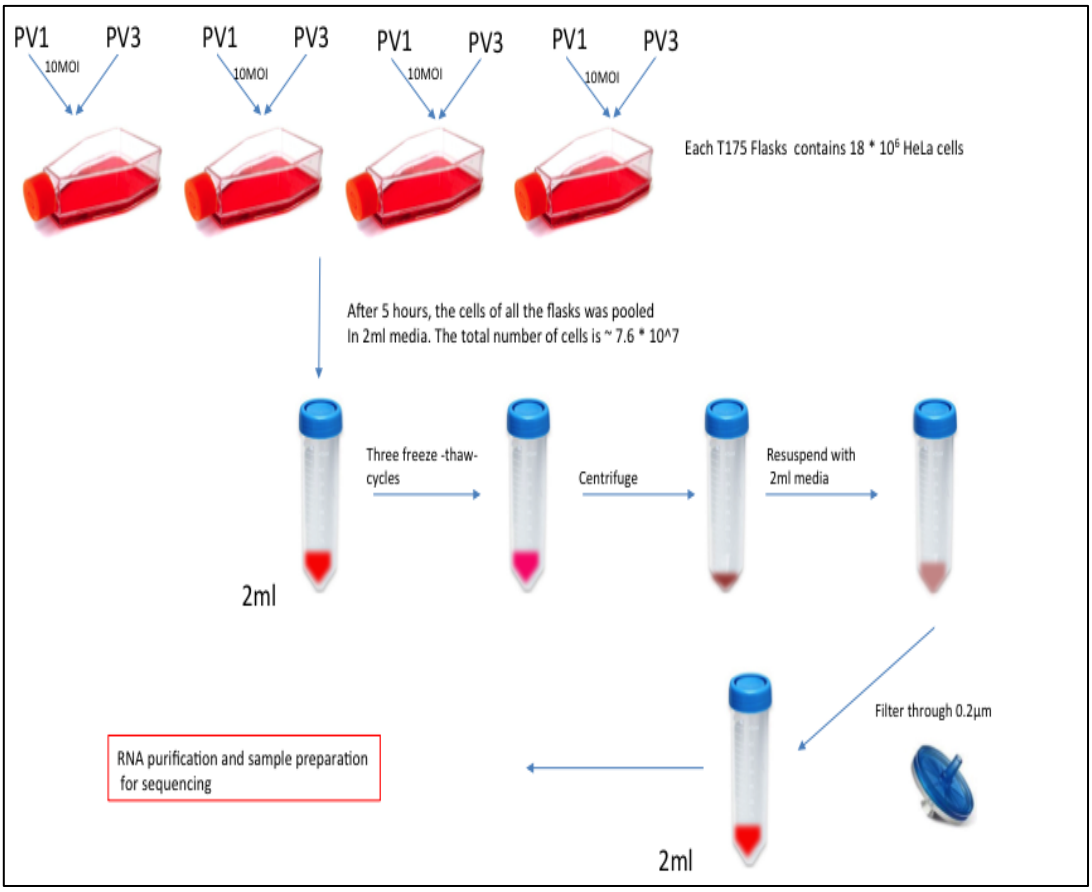


Figure 4-2 Schematic drawing of the virus co-infection experiment
The co-infection experiment was carried out in four T175 in parallel, and after 5 hours the infected cells were pooled together in 2ml media. The pool of the cells was lysed to extract the virus and the debris was discarded. The supernatant contained the viruses and the cellular molecules were filtered through 0.2µm filters. The resulted 2ml filtrate was used for RNA extraction, subsequent RT-PCR and sequencing

4.3.1 Optimising the annealing temperature of the specific-recombinants primers

To guarantee the efficient and specific amplification of recombinants, the annealing temperature for primP2-R_LD_M was optimised using a temperature gradients PCR. The ideal situation would be the ability of primP2-R_LD_M to amplify a product of 1313bp when the PCR template was a recombinant molecule (positive control) and no products when the template was either PV1 or PV3 genome (negative control). A plasmid that contained a molecular clone of an imprecise-insertion recombinant, which contained 249 nts insertion, was used as a PCR template. This was constructed by Dr. Jonathon Cook and based on an isolate from the recP2-R_LD_M found by Lowry *et al.* (henceforward known as construct JC105B). It is worth highlighting that as this molecule included 249 extra nucleotides, the expected band would become 1565bp rather than 1313bp.

At the beginning of this project, PCR gradient reactions were performed between 54.1ºC and 61.1ºC, an example of which is given in Figure 4-4. However, as most of the tested primers showed the best performance between 57.1ºC and 60.1ºC, based on the intensity of the specific band and the absence of non-specific products, the range 57.1ºC to 60.1ºC was used to optimise the primers at a later stage of the project. The primP2-R_LD_M was found to amplify correctly within this range, without any non-specific priming on the parental genomes (Figure 4-5). The Primers generated a band of the expected size when ~0.3 ng of JC105B DNA was included in the reaction, and no products were observed when either full-length PV1 pPV1FLC plasmid or PV3 full-length pT7FLC with the same concentration was used as a template (henceforth referred to as PV1 and PV3 plasmids). Due to the lack of differences in amplification between the temperatures, any of them could be used as an optimised annealing temperature. Nevertheless, for the sake of consistency, 58.1ºC was used for primP2-R_LD_M throughout this study. This was repeated several times before being adopted.

4.3.2 Measuring the PCR sensitivity and testing for artefacts using plasmid DNA

Having determined the annealing temperature, which showed a specific amplification for recombinants molecules at 58.1ºC, it was important to test the efficiency and the sensitivity using this temperature under the same PCR conditions (see previous section). Therefore, three different conditions were analysed:

1. The presence of artefacts (false positive recombinants).
2. The sensitivity of the PCR reaction in amplifying recombinants alone.
3. The sensitivity of the PCR reaction in amplifying recombinants in the presence of the parental genome.

All the above-mentioned cases were optimised by using the PV1, PV3, and JC105B plasmids. In the first case, the formation of artefacts was tested. It is known that when PCR amplification is carried out using a mixture of two or more non-identical but similar genotype target sequences as a template, it can lead to a formation of chimeric artifacts by template switching (Lahr and Katz, 2009). This is important, because the designed experiment (Figure 4-2) would generate samples contain both parental and recombinant genomes. Therefore, the formation of these chimeric molecules was possible, which – if occurred - could have underrepresented the presence of recombinants formed between the viruses inside the cells. To test this, a PCR reaction containing primP2-R_LD_M with ~0.3ng of an equimolar mixture composed of PV1 and PV3 plasmids was prepared as a template.

The second case, because recombination is a rare event that can exist at a low frequency, it was important to measure the analytical sensitivity of the PCR assay *i.e.* the lowest number of DNA copies that the assay could amplify. Determining its sensitivity would help in deciding the scale of the experiment later. This was measured by carrying out a serial 10-fold dilution of the JC105B plasmid, which contained a full-length imprecise-insertion recombination with 249 nts insertion. The dilution series – diluted in nuclease free water - started with ~10^7 copies of JC105B and was diluted down to reach ~10 copies at the final dilution.

The third case was related to the second one, but differed in that it measured the PCR assay sensitivity of detecting recombinants in the presence of parental genomes. The importance of this stems from the fact that the primers could be used up by binding

to one of the parental genomes before reaching the targeted recombinants. Thus, adjusting the quantity and the specificity of the primers was crucial in this context. This was tested by spiking a mix of ~10^7 copies of each of the two parental virus plasmids with each of the dilutions in the JC105B dilution series.

The prepared samples of all of the cases above were PCR amplified in parallel under the same conditions (see previous sectionS). When a mixture of both viruses was used as a template, no PCR product was observed, which showed that no artefacts were detected (Figure 4-6 A). On the contrary, a band of the expected size was formed when the template used was JC105B (Figure 4-6 B, '+'), which confirmed that the absence of artefacts in the previous sample was not due to a technical error.

When the concentration of the template was serially diluted, the PCR assay demonstrated high sensitivity as it could detect down to ~10 copies of the DNA (Figure 4-6, B). The intensity of the bands started to decrease gradually when ~1000 copies of DNA were used as a template, and was barely seen with ~10 DNA copies. However, the fact that the band could still be seen on the gel suggested a successful detection. Comparably, when JC105B was spiked into a fixed concentration of parental viruses DNA, the PCR assay could detect down to the presence of 100 DNA copies (Figure 4-6, C). This suggested a 10-fold decrease in the PCR sensitivity of detection in the presence of parental viruses. In both cases, small visible bands could be observed at the bottom of the gel. These could be clearly seen when ~10^4 copies of DNA were used as a template and their intensity increased with the decrease of the templates' copy number (Figure 4-6, B & C). Based on their small size (below the smallest band in the ladder) these bands could have possibly resulted from excess primers left over after the PCR amplification. Additionally, they did not appear when the DNA copy number was higher than ~10^4 copies. The rationale is that the primer concentration had become higher than the targeted template concentration. On the other hand, the fact that artefacts did not appear on the gel did not rule out their presence at <10 copies. Nevertheless, this could be tackled during the NGS analysis by running a control sample that contained a mixture of both viruses (see section 5.4).

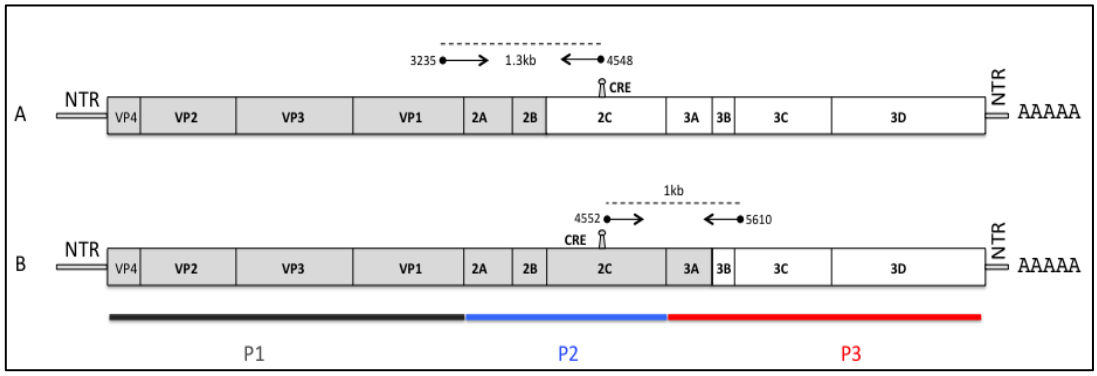


Figure 4-3 Poliovirus genome map with the PCR targeted regions indicated
Two possible recombinants genomes are illustrated to highlight the locations of the targeted regions considered in this project. The grey colour refers to the PV3 genome and the white colour to the PV1. The dashed line denotes the region of interest. The small arrows represent the primers with their 3' position indicated. The black, blue, and red lines at the bottom refer to the P1, P2, and P3 regions of polioviruses genome respectively. The non-translated regions are represented by NTR. The CRE and the coding regions are indicated. A) Example recombinant demonstrates the primP2-R_LD_M. B) Example recombinant demonstrates the primP3-R_LD_M. The locations were located on the genome drawings using poliovirus complete genome with NCBI Reference Sequence: NC_002058.3. The CRE extends from nucleotide 4435 to 4495 based on the PV3 genome (Goodfellow et al., 2000)

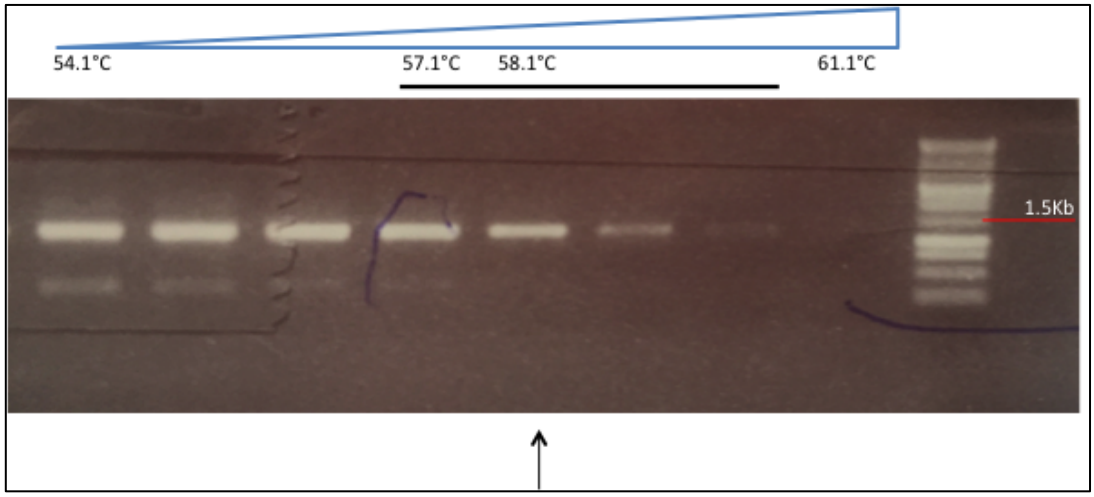


Figure 4-4 An example of PCR annealing temperature optimisation
The figure shows an optimisation of the annealing temperature using another pair of primers than primP2-R_LD_M earlier on in the project. The amplicons in the figure belong to the construct JC105B amplified with a pair of primers that produced a product of ~1100bp. However, as JC105B contains an insertion of 249 nts, the bands appear at the size of ~1349bp. The triangle refers to a gradual increase in the annealing temperature. The black arrow shows the lane that corresponds to the optimal amplification where the band intensity is the highest in the absence of the non-specific band. Most of the tested primers showed their best performance between 57.1ºC and 60.1ºC (the black horizontal line). Therefore this range was adopted for the rest of the optimizations carried out throughout the project.

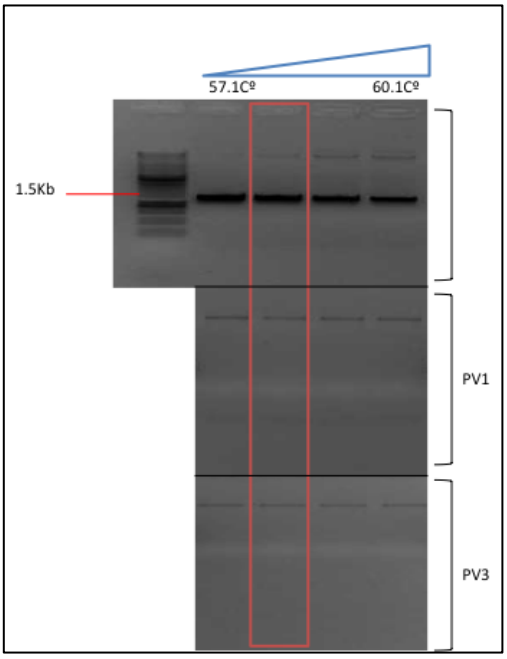


Figure 4-5 The specificity of the detection at different annealing temperature using primP2-R_LD_M (primers within the P2 region)
The construct JC105B and the plasmids of the parental viruses were amplified individually using primP2-R_LD_M at different annealing temperatures starting from 57ºC to 60.1ºC. The PCR products run on 1% agarose gel and were stained with ethidium bromide. The triangle refers to a gradual increase in the annealing temperature. When JC105B was used as a template, the band of the expected size was observed in all the considered annealing temperatures (the gel on the top). When either PV1 (the gel in the middle) or PV3 (the gel at the bottom) were used as templates no products were observed at any annealing temperature. The red rectangle indicates the lane where the annealing temperature was set to 58.1, the one that was chosen to be used throughout this project.

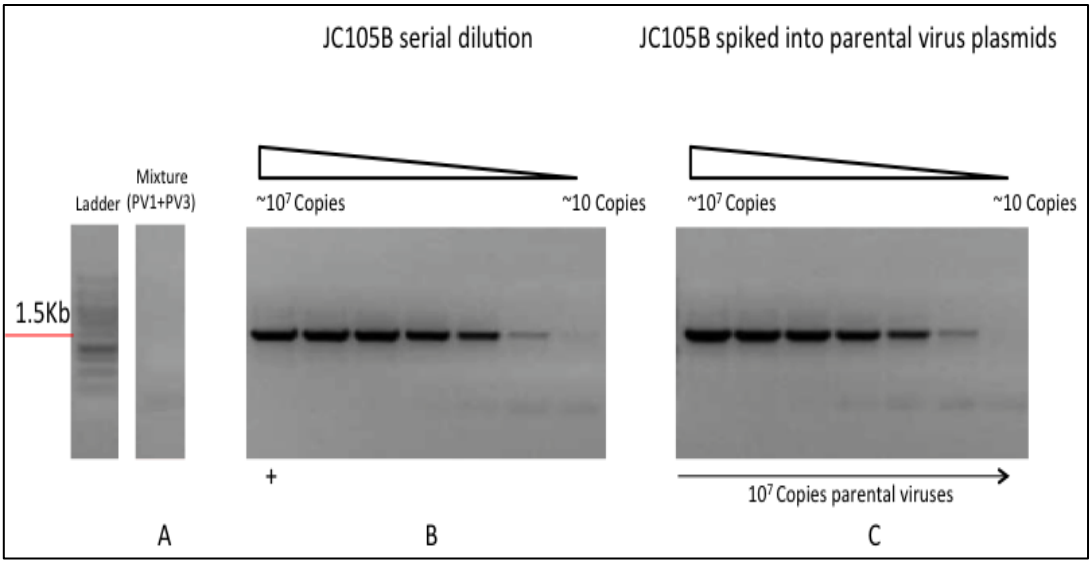


Figure 4-6 Sensitivity and artefacts optimisation of PCR with primP2-R_LD_M (primers within the P2 region)
Products from PCR reactions were separated by agarose gel electrophoresis, stained with ethidium bromide, visualised on a transilllluminator and photographed. (A) Shows negative control, a mixture of PV1 and PV3 plasmids, (B) shows PCR products amplified from the JC105B dilutions and C) shows PCR products from JC105B dilutions spiked into a fixed concentration of parental viruses. The sign + represents the line of the positive control, before starting the dilution. The amplification of the JC105B in the '+' lane is the same amplification of the first dilution in the series, hence it appears in one lane.

4.3.3 cDNA synthesis sensitivity

Based on the PCR sensitivity results, the samples generated by the co-infection experiment (Figure 4-2) should contain at least 100 recombinant molecules to be detected in the PCR assay in the presence of parental genomes. However, as these samples would contain RNA, which had to be reverse transcribed into cDNA before PCR amplified, the sensitivity of cDNA synthesis should be at the same level or higher so as to synthesise enough cDNA to be detectable by the PCR reactions. Thus, it was important to measure the cDNA synthesis sensitivity. The detection sensitivity of the cDNA synthesis was performed at an early stage in the project by using the PV3 full-length pT7FLC. The plasmid was linearized with *Sal*I, followed by RNA transcription using T7 polymerase. The resulting reactions that contained RNA were DNAse treated by TURBO DNAse to remove any residual cDNA. Subsequently, the RNA was diluted in serial 10-fold increments starting from ~$5*10^5$ RNA copies moving down to ~5 RNA copies, and cDNA was synthesised using Superscript III Reverse Transcriptase (RT) and oligo dT primers (see section 2.2.9). Eventually, 5µl of the resulted cDNA were PCR amplified using PV3-specific primers (PV3-3235F and PV3-4521R) that generated a product of 1.3kb (Table 2-3).

When the plasmid was omitted from the RNA transcription reaction (negative control), the PCR amplification did not produce any product that reflected the absence of contamination (Figure 4-7, A). On the other hand, the PCR amplification demonstrated a high sensitivity by being able to detect as low as ~5 copies of cDNA synthesised using RT (Figure 4-7, B). This means that the cDNA synthesis reaction could reverse transcribe down to ~5 copies of RNA. To rule out that these products resulted from amplifying leftover DNA from the RNA transcription reaction, a negative (no-Reverse Transcriptase) reaction was considered for each dilution from $5*10^5$ to 5 DNA copies (Figure 4-7, C). Clearly, no bands of the expected size were observed. This, together with the fact that DNAse treatment was carried out, confirms that PCR products of the dilution series samples resulted from RNA that had been reverse transcribed (Figure 4-7, B). Finally, the small visible bands observed at the bottom of the gel most probably represented leftover reactants.

4.3.4 Quantitative estimation of the recombinants generated by the experiment

One of the key issues in the proposed method was to generate a sufficient amount of recombinants within the detectable limit of the subsequent cDNA synthesis and the PCR amplification. Based on the sensitivity results (see sections 4.3.1, 4.3.2), which demonstrated that the PCR assay could detect down to 100 recombinant copies in the presence of parental viruses, a large-scale experiment was decided to ensure obtaining more than 100 recombinants in the final samples. This section explains the basic calculations used to estimate the number of recombinants molecules in the samples.

Four T175 flasks, each containing $18.6*10^6$ cells, were co-infected with poliovirus type 1 and type 3. To assure that the vast majority of the cells would be infected with at least two viruses the infection was carried out at MOI of 10 using 0.2ml of each parental virus stock with concentration of 10^9 Plaque Forming Unit (PFU)/ml ($0.2 * 10^9 / 18.6 * 10^6 = \sim 10$ MOI), which means that there were 10^9 PFU * 0.2ml = ~ $2*10^8$ viruses from each type. In four flasks there were $(2 * 10^8 * 2) * 4 = 1.6 * 10^9$ viruses from both types. After five hours, as measured by the plaque assay, the virus titer would increase by around ~ 4 logs (Figure 4-8), which translates into $1.6 * 10^{13}$ viruses. These viruses were then collected in 2ml media, in which – based on the recombination ration 1/100000 - the number of recombinants was estimated to be $1.6 * 10^8$. From this, only 140µl was used to extract the viral RNA, which means the number of recombinants would become $1.1 * 10^7$. In 40µg total RNA extracted, the recombinants would constitute 0.1ng from the total. From the total RNA, 3µg was used to synthesize the cDNA, which would have 4000 copies of recombinants. The copy number was calculated by the previously mentioned calculator (http://cels.uri.edu/gsc/cndna.html). However the calculation above was based on the assumption that no loss happened during the collecting and the extracting steps.

Because it is hard – within the context of this project – to be certain of how many particles were infectious, or how many recombinants were losses, or how many particles invaded each cell this was not explored any further.

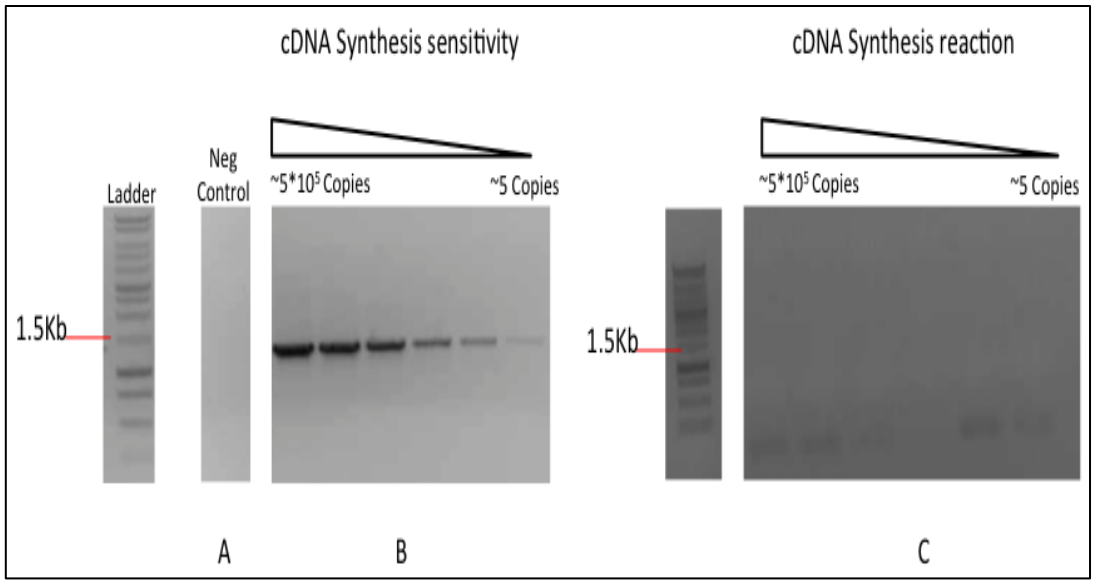


Figure 4-7 cDNA synthesis sensitivity (primP2-R_LD_M)
Products from PCR reactions were separated by agarose gel electrophoresis, stained with ethidium bromide, visualised on a transilluminator and photographed. PV3 full-length pT7FLC was used with PV3-specific primers A) Negative (no-plasmid) control. B) PCR amplification of the RNA dilutions that had reverse transcribed. C) PCR amplification of the negative (no-RT) reactions.

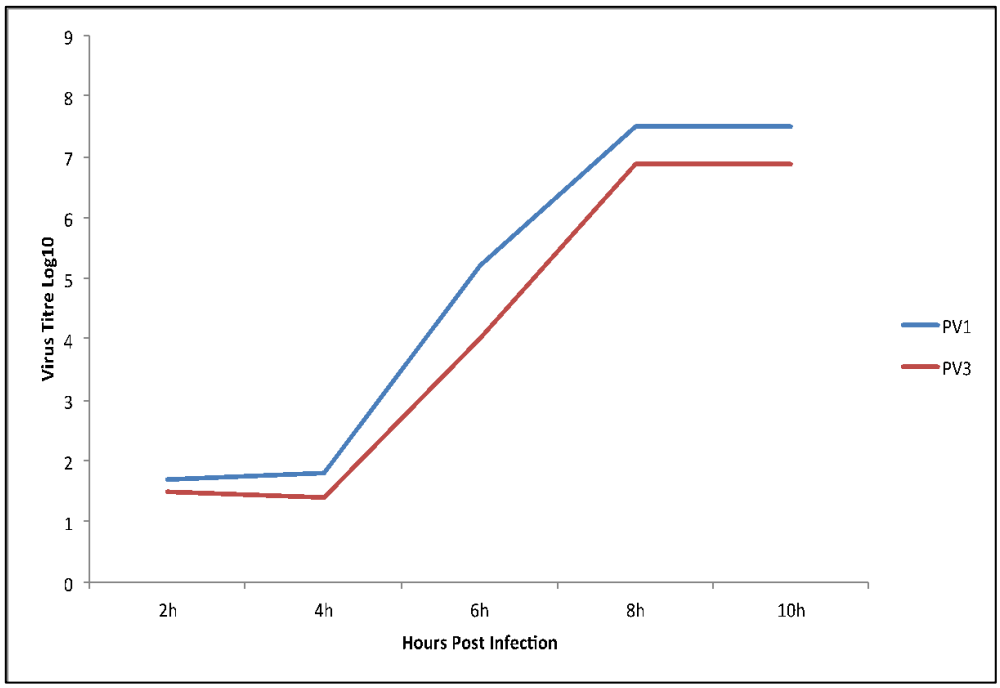


Figure 4-8 Single-step growth analysis over 12 hours of parental viruses
Comparison between the growth characteristics of parental viruses (PV3 and PV1) was measured by the plaque assay. The y-axis refers to the virus titer (pfu/ml) and the x-axis to time post infection.

4.4 Amplification and Sanger sequencing of the recombinant amplicons

4.4.1 Amplification of recombinants within the P2 region of the virus (recP2-R_LD_M)

The next step was to amplify, sequence and analyse the recP2-R_LD_M which resulted from the co-infection of HeLa cells with wild-type poliovirus type 1 and type 3 (Figure 4-2). The four T175 flasks, in which HeLa cells were co-infected with both viruses, were pooled together in 2ml media (co-infection-sample), the viral RNA was then extracted, and the cDNA was synthesised using Superscript III and oligo dT primers, followed by PCR amplification. In addition to that, two "no co-infection" control samples (PV1-sample and PV3-sample), where HeLa cells were infected with each virus in separate cultures, and one mock infection control (mock-sample) were generated in the same way. In order to verify the recombination products, three different controls were analysed in parallel.

Firstly, to test if the infection experiment and PCR amplification had worked as expected, PV1-specific and PV3-specific primers (Table 2-3) were used to amplify cDNA from PV1-sample and PV3-sample respectively (Figure 4-9, A, lane 1 & 2). Products of the expected size (1.3kb) were observed, which indicated that PV1 and PV3 viruses existed and the targeted genomes were successfully amplified. The identity of each product was further verified by Sanger sequencing to ensure that they were not derived from the same virus (data not shown).

Secondly, to verify the absence of the artefacts an equimolar mixture of viral RNA was prepared and used to make the cDNA. Subsequently the cDNA of the mixture was used as a template in a PCR reaction using primP2-R_LD_M (Figure 4-9, A, lane 3). In agreement with the optimisation results (Figure 4-6), no products of any size could be seen, which ensured the absence of artefacts. It was important in this control to mix the viruses before synthesising the cDNA, as there was a possibility that during the cDNA synthesis, rather than during the PCR reaction, artefacts could have formed.

Thirdly, PV3-specific and PV1-specific primers were used to verify the presence of both viruses in the co-infection-sample (Figure 4-9, B, lane 4 & 5). The appearance of both products with the expected size proved that both viruses existed in the sample and took part in the infection. Eventually, primP2-R_LD_M was used to amplify recombinants in the co-infection sample. (Figure 4-9, B, lane 6). A band of the expected size with low intensity was observed. Depending on the control results, this band was interpreted as representing one ore more recombination products. To test whether the amplification of recombinants from this sample was reproducible, another cDNA were synthesised from the co-infection-sample, PCR amplified, and run simultaneously on the gel (Figure 4-9, B, lane 7); here the same product was observed. To further analyse these products, the camera was zoomed-in and the gel was photographed again (Figure 4-9, C, lane 6'& 7'). This revealed an extra ~1kb band with a smear above and below the major band, which seemed to be a reflection of imprecise recombinants (deletion and insertion).

4.4.2 Increasing the yield of recombinants' amplicons within the P2 region by pooling 8 PCR reactions

To this end, as the main aim of this study was to identify and characterise recombinants by next-generation sequencing, it was necessary to confirm the recombination products in Figure 4-9 by cloning and Sanger sequencing. The main goal of this was to test whether the amplified recombinants were derived from a diverse recombination population, thus eliminating the possibility of ending up sequencing one or two amplified molecules.

The first preparation step was to increase the yield of amplified recombinants, which would provide a sufficient amount to guarantee an efficient cloning and subsequent sequencings. To do this, 8 PCR reactions using the cDNA from the co-infection-sample were performed in parallel, in which primP2-R_LD_M were used to amplify recP2-R_LD_M. After amplification, all the reactions were pooled in one tube and treated as one sample, the PCR products were purified and separated by agarose gel. Concurrently, two control samples were considered and prepared in the same way. The first one was to test for any contamination that may have developed during the process; the mock-sample was used for this purpose (Figure 4-10, lane 1). The second was to test for any artefacts formation; as the scale of the PCR reaction was

enlarged, thus it was more likely to see artefacts than in the case of one PCR reaction if there were any (Figure 4-10, lane 2). Nevertheless, neither artefacts nor contamination were observed. On the contrary, a band with high intensity of the expected size was observed in the co-infection-sample (Figure 4-10, lane 3), indicating that recombinants were successfully amplified with a high yield.

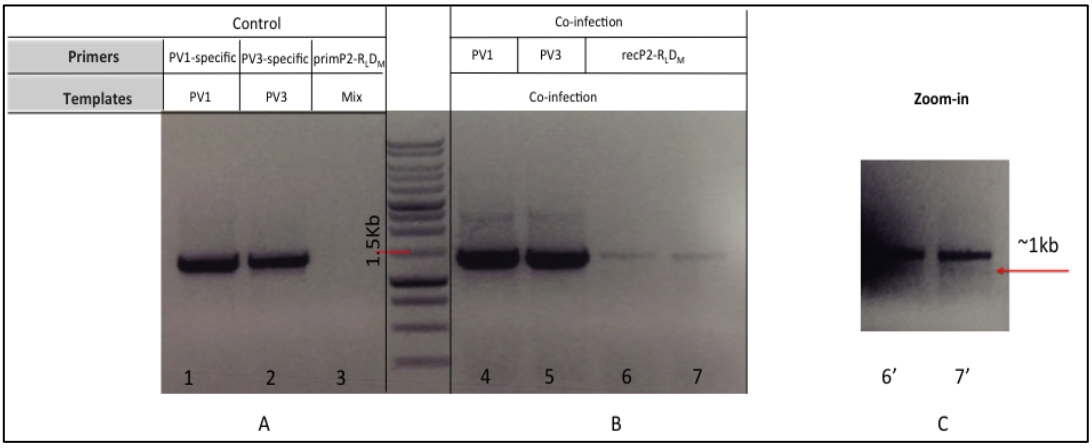


Figure 4-9 Recombinants amplification within the region of P2 (recP2-R_MD_L) from the Co-infection sample
The viral RNA - extracted from the PV1-sample, PV3-sample, and the co-infection-sample – was cDNA synthesised and amplified with different primers before running on 1% agarose gel and being stained with ethidium bromide. The 'Primers' in the table above the gel refers to the primer pairs used while the 'Templates' indicates the cDNA used. Mix = RNA mixture of both viruses used to synthesise the cDNA, which then amplified A) Control reactions were used to confirm the absence of any cross-reactivity, and to ensure that the infection had worked. B) Different primers were used to amplify the cDNA from the co-infection-samples C) Zoom in on Lane 6 and 7 to show the presence of an extra band and smears within the recombination PCR products.

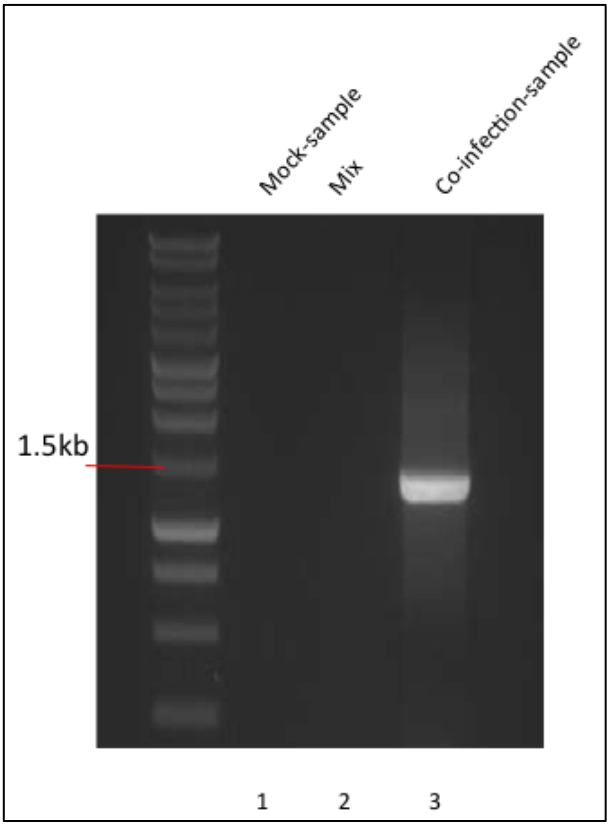


Figure 4-10 Increasing the yield of recP2-R_MD_L by pooling 8 PCR reactions
To increase the yield of the PCR amplification, the products of 8 reactions were pooled in one tube, purified and then screened on agarose gel, stained with ethidium bromide and photographed. Lane 1 and 2 are negative controls. Lane 3 is the co-infection sample.

4.4.3 Cloning and Sanger sequencing of recombinants within the P2 region (recP2-R_LD_M)

From the recombinants products on the gel (Figure 4-10), three separate regions were excised, purified, and cloned into pGEM-T Easy Vector separately. The region above the major band was named 'insertion' indicating the imprecise-insertion recombinants. The region of the major band (1.3kb) was named 'precise', signifying precise recombinants with no insertion or deletion. The region below the major band was named deletion representing the imprecise-deletion recombinants (Figure 4-11 A). The cloned vectors of the three regions were transformed into competent *Escherichia coli,* plated onto the LB agar plate and incubated overnight at 37ºC. The next day, several clones from each region were picked, lysed in distilled water, PCR amplified using primP2-R_LD_M, and photographed. This process was repeated several times to pick as many colonies as the time allowed.

Several clones were successfully isolated from the 'precise' and 'deletion' regions. In contrast, successful cloning and amplification was limited regarding the 'insertion' region, possibly due to the length of recombinants in this region, which might lessen the efficiency of cloning. This could explain the fact that the only successfully isolated clone with insertion had only 2 extra nucleotides. Examples of the clones that were successfully amplified from the three regions are shown in Figure 4-11 B. The insertion and precise products appeared to be the same length (Figure 4-11, lane 2'and 3'), as they only differed by 2 nucleotides and were both clearly shorter than the control JC105B. Right to the ladder, the shorter band (~700 nts) reflected a recombinant that had lost part of its genome (Figure 4-11, lane 4'). These PCR amplicons, in addition to another 12 amplicons, were prepared for Sanger sequencing (see section 2.2.7) and sequenced by the University of Warwick Genomics Facility.

The resulted sequences were analysed by ViReMa – the algorithm described in section 3.5 – to detect the junction locations and were then visualised on the Parallel Coordinates, a visualisation tool (see section 2.8.14) that was developed for this purpose (Figure 4-12). The lines within the map show the connection between the location from where the RdRp had disassociated from the donor and re-associated to

the acceptor. The lines are horizontal (slop = 0) if the recombinants did not contain any insertion or deletion.

From the 15 analysed recombinants, 11 were precise recombinants with 6 unique recombination junctions, as some of them were found in multiple copies. Precise recombinants are, by definition, in-frame; therefore all of these recombinants were represented by horizontal blue lines (slope = 0). However, if insertion or deletion was involved, the slope would change, unless the inserted or deleted fragments were ~10 nts length, in which case the line would appear horizontal as with the precise recombinant line. In fact, this is the case of the isolated imprecise-insertion recombinant, which contains insertion with 2 nts length (the line is marked with a star in Figure 4-12). Although the line looks horizontal in the map its colour is red, which makes it easier to tell that it is not precise. If the length of the insertion fragment increased, the lines' slope would then increase; as a result, the lines would point up towards the 3' end of the recipient genome. On the contrary, the lines would point down towards the 5' end of the recipient genome if the recombinants had part of their genomes deleted. This was the case for the three analysed imprecise-deletion recombinants, two of which were found to occur within the *2C*-coding region, and one within the *2A*-coding region.

Interestingly, the lengths of deleted fragments were found to vary. For example the in-frame imprecise-deletion recombinant occurred at the *2C*-encoding region in the donor strand that had resulted from the RdRp switching to the *2A*-encoding region in the recipient strand. This was represented by a fragment of ~1100 nts length being deleted from the recombinant sequence and, thus, represented by a line being pointed down towards the 5' end of the acceptor (negative slope). However, the slope was increased for the line which represents imprecise-deletion recombinants occurring between the *2A*- and *VP1*-encoding regions as the deleted fragment length decreased to ~89 nts (Figure 4-12). These findings show that the isolated recombinants were diverse and suggest that they were not derived from over-amplification of one recombinant molecule.

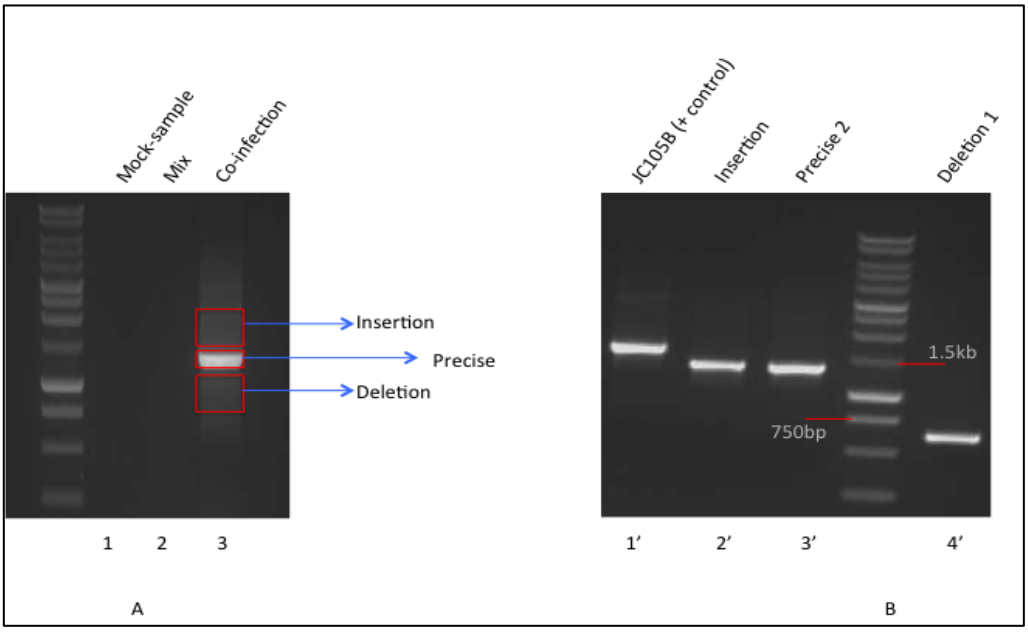


Figure 4-11 Gel extraction of the PCR products that corresponded to precise and imprecise recombination within the P2 region (recP2-R_LD_M)
A) A pool of 8 PCR products was amplified by primP2-R_LD_M run on 1% agarose gel and stained with ethidium bromide. Mock-sample and Mix are negative controls. The red rectangles demonstrate the regions that were excised from the gel for the cloning step, while the blue arrows indicate the names (types) of recombination of each region. B) PCR products of three colonies amplified by primP2-R_LD_M represent precise and imprecise recombinants. Lane 1' is the amplification product of JC105B (positive control).

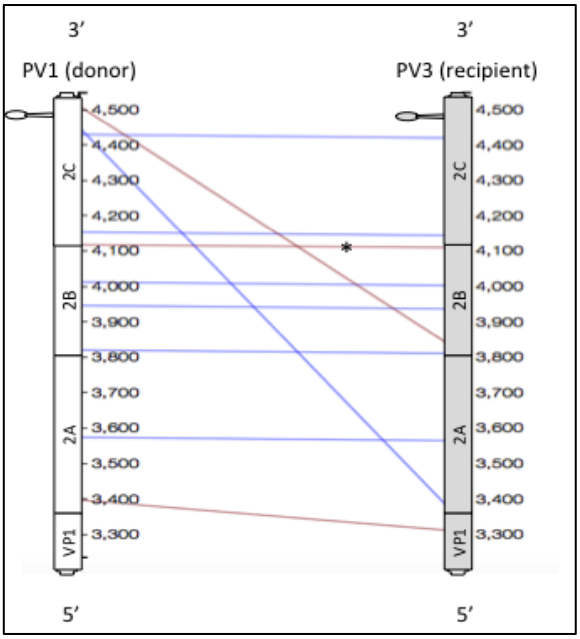


Figure 4-12 Parallel Coordinates map of recombinants amplified from the P2 region (recP2-R_LD_M)
The amplified recombinants were cloned, sequenced, and analysed by the algorithm ViReMa (see section 3.5). The detected junction locations were visualized using the Parallel Coordinates tool, which was developed for this project. The PV1 genome (the donor strand) is in white on the left side. The recipient genome is on the right side. The genomes are represented by cartoon drawings referring to the regions of interest (within the P2 region) with the coding regions indicated. The lines between the genomes represent recombinants and connect between the junction locations in both genomes. The red lines denote out-of-frame recombinants, and blue lines indicate in-frame recombinants. The slope of the line reflects the size of inserted or deleted fragments, while the direction of the line refers to the type of recombinants (horizontal line = precise, negative-slope line = deletion, positive-slope line = insertion). The 10 analysed recombinants from recP2-R_LD_M are demonstrated. The star highlights the imprecise-insertion recombinant.

4.4.4 Amplification of the reciprocal recombinants' population within the P2 region (recP2-R_MD_L)

The analysis results of recP2-R_LD_M populations made it worth expanding the search for other recombinants populations. From the same P2 region, recombinants that happened in the opposite direction (reciprocal) were analysed. In this population, the PV3 virus would become the donor and the PV1 the recipient, hence the name (recP2-R_MD_L). The primers used to amplify this recombinant population were designed from the same area in the genome (PV1-3241F and PV3-4521R and Table 2-3). Using the same annealing temperature (58.1ºC) and the same PCR conditions, the primers worked successfully in amplifying the recombinants without any evidence of artefacts when tested with the RNA mixture of both viruses (Figure 4-13 A).

The same strategy was used to analyse this population. The primers were used to amplify 8 PCR reactions in parallel. Subsequently, they were pooled in one tube, separated by agarose gel, cloned into pGEM-T Easy Vector, sequenced, and visualised on the Parallel Coordinates (Figure 4-13 B). From the total of 7 recombinants sequenced, 6 were precise recombinants with 5 unique junctions. The remaining recombinant was imprecise-deletion out-of-frame with ~20 nts deletion. Of the five recombinants, two were found in the *2C*-coding region, two in the *2B*-coding region, and one in the *2A*-coding region. On the other hand the imprecise-deletion was located at the beginning of the *2C*-coding region. No colonies were isolated containing plasmids carrying imprecise-insertion recombinants.

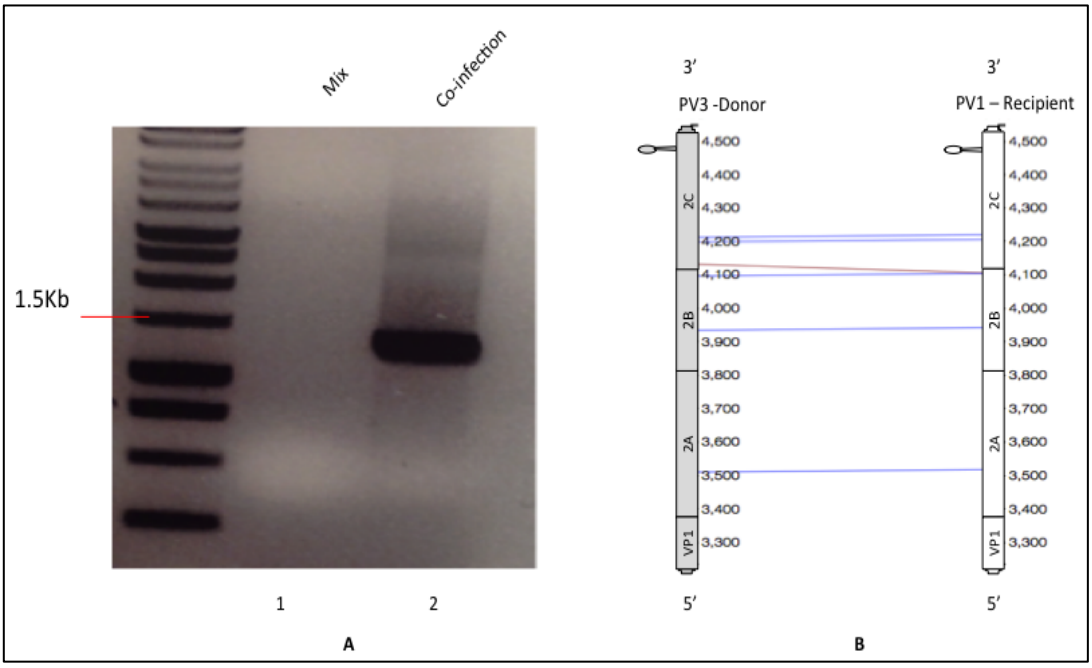


Figure 4-13 recP2-R_MD_L; recombinants' population (Reciprocal population)
A) PCR products amplified with primP2-R_MD_L, screened on 1% agarose gel, and stained with ethidium bromide. Mix = cDNA from RNA mixture of both viruses. B) Parallel Coordinates visualisation of recombinants isolated from recP2-R_LD_M population.

4.4.5 Amplification of the recombinant populations within the P2-P3 region (recP3-R_LD_M and recP3-R_MD_L)

After succeeding in isolating several different recombinants from the P2 region, the search process was further expanded to the region downstream of the CRE until the first part of the *3C*-coding region, which encompasses the *3A*- and the *3B*-coding sequences from the P3 region in the genome (Figure 4-3 B). Therefore the P2 was replaced by P3 in the names. Primers for recombinants populations were designed in the same way in the P2 region (see section 4.1.1); the populations of the two possible directions (recP3-R_LD_M & recP3-R_MD_L) were targeted. When testing the designed premiers (PV3-4552F - PV15610R & PV1-4579F - PV3-5647R, Table 2-3) on the co-infection sample, a product of the expected size (~1.1kb) was observed for both directions (Figure 4-14, lane 2, 4), which suggested that recombination occurred in this region of the genome. Nonetheless, when using an equimolar mixture of viral RNA as a control, a band of the expected size was observed (Figure 4-14, lane 1, 3), which indicated a non-specific priming event had taken place in the reaction. Additionally, an extra smaller band was observed in the mixture reaction (~250 nts) using primP3-R_MD_L (Figure 4-14, lane 1 with red rectangle). Sequencing these products revealed that they were derived from either PV3 or PV1 genomes, which suggested a mispriming event rather than a formation of artefacts. Probably, this non-specific priming happened because of the high similarities in this region (~85%) represented by long identical stretches (~25-30 nts) between the two genomes at this part of the genome.

Having a defined recombinant construct (JC105B) in hand helped to optimise the primers for the P2 region. However, this was not the case for this region of the genome, as no such a recombinant control was available to optimise the reaction. Therefore, advantage was taken of the co-infection-sample, which was used as a positive control. The yield of amplifying the viral RNA mixture of both parents by primP3-R_MD_L and primP3-R_LD_M (Figure 4-14, Lane1 & 3) appeared to be lower than the yield of the PCR products which resulted from amplifying the co-infection-sample with the same primers (Figure 4-14, Lane2 & 4). This observation is evidence that both specific and non-specific priming happened in the co-infection-

sample. Accordingly, the annealing temperature was increased by one degree to be 59.1ºC and 8 PCR amplifications were performed again and screened on agarose gel. It was reasoned that increasing the annealing temperature would decrease the non-specific priming within the co-infection-sample and provide purer recombinants amplicons for cloning. Increasing the annealing temperature had, in fact, succeeded in lowering the amount of the non-specific products (Figure 4-15 A).

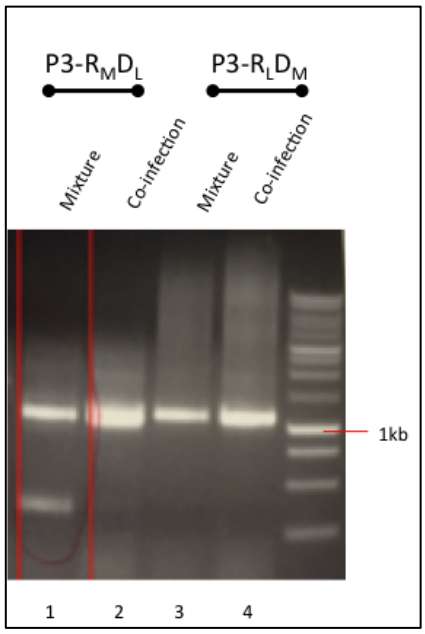


Figure 4-14 Optimisation of recombinants' amplification within the P2-P3 region (recP3-R$_M$D$_L$ and recP3-R$_L$D$_M$)
PCR products were amplified with primP3-R$_M$D$_L$ and primP3-R$_L$D$_M$ at 58.1ºC annealing temperature, separated by 1% agarose gel, and stained with ethidium bromide. Mixture = cDNA from RNA mixture of both viruses. The red rectangle highlights the lane where more than one non-specific product was observed. The products in lane 1 and 3 represent non-specific products which resulted from amplifying the viral RNA mixture. The remaining lanes represent amplification from the co-infection-sample. The lines above the gel categorise the lanes into their corresponding population.

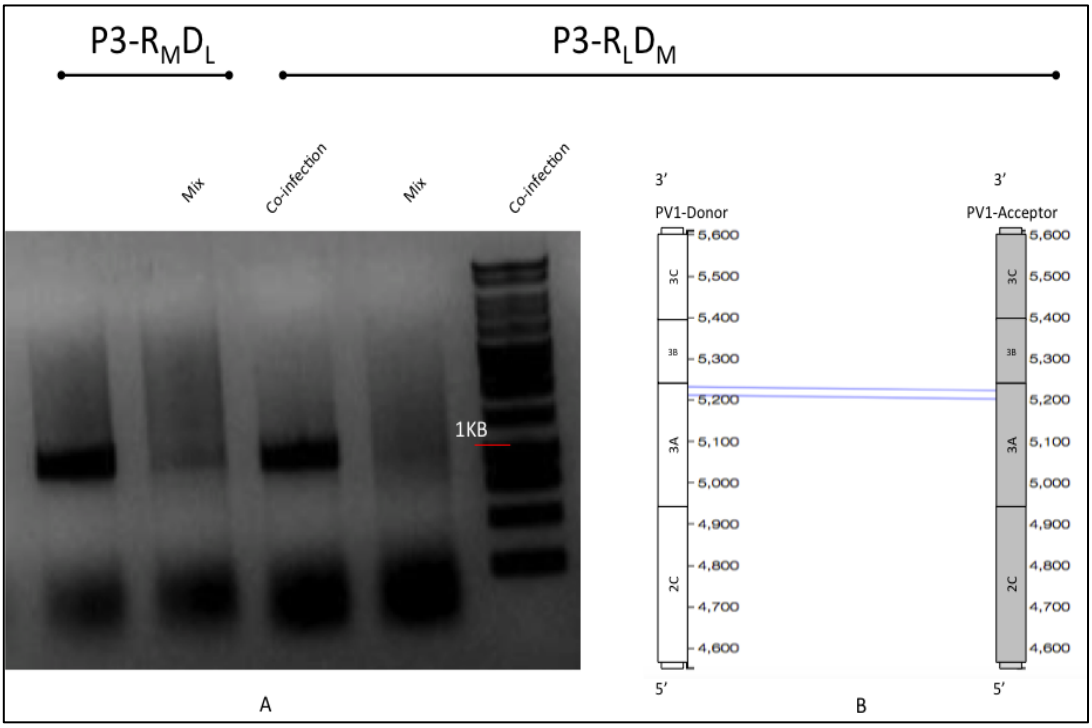


Figure 4-15 Amplification of recombinants within the P2-P3 region (recP3-R_MD_L and recP3-R_LD_M)
A) PCR products amplified with primP3-R_MD_L and primP3-R_LD_M at 59.1ºC annealing temperature, run on 1% agarose gel, and stained with ethidium bromide. Mix = cDNA from RNA mixture of both viruses. B) Parallel Coordinates visualization of recombinants which wer amplified from the recP3-R_LD_M. The left line above the gel categorises the lanes where primP3-R_MD_L was used. The right line is categorising the lanes wherein primP3-R_LD_M was used and correlating them to the Parallel Coordinates visualisation map.

Although non-specific bands were observed at the bottom of the gel (Figure 4-15 A), the recombinants amplicons of the expected size was now much clearer for analysis. These small products could have been formed due to a cross reaction between the primers (cross-dimers), which is likely to have happened in this part of the genome where the similarity is high. Nevertheless, these products were removed when preparing the samples for NGS.

As the time started to become very limited, only three molecules from recP3-R_LD_M were cloned and sequenced to verify the presence of recombinants in this region. Of those, two unique precise junctions were identified at the end of region encoding 3A protein (Figure 4-15 B).

4.5 Preparing samples for next generation sequencing (NGS)

Based on the success in identifying recombinants from different locations and of different types, the use of next-generation sequencing was valid and justified to identify a wider range of recombinants. In order to achieve this, 7 major samples were prepared by purifying the recombinants amplicons and sent off to GATC Ltd. (www.gatc-biotech.com) for sequencing by Illumina MiSeq. Four of them represented the four recombinants' populations described in this chapter, and two samples were negative controls composed of viral RNA mixture treated with the same primers used to amplify the co-infection. The remaining sample was a known concentration of recombinant JC105B spiked into known concentrations of the parental viruses to establish a method to quantify recombinants based on the NGS reads.

All the samples were prepared to meet the requirements of the company, in terms of quantity and quality. The agreement was to fragment the amplicons in each sample (~1.3kb & ~1kb) into ~550 nts and sequence them with 300 nts paired-end reads by Illumina technology. The company guaranteed the feasibility of the project, and assured us that they would provide us with high-quality results.

Unfortunately, after receiving the samples, the company – after much further delay – came back to us indicating that they could not complete the sequencing due to the

way the samples had been prepared. GATC Ltd. subsequently admitted liability for providing incorrect information and for destroying the material submitted to them.

As the time had become very narrow, the only solution was to repeat the experiment or part of it in order to generate the NGS dataset. My supervisor Professor David Evans contacted Dr. Andrew Macadam at National Institute for Biological Standards and Control (NIBSC) and arranged a two-week visit for me to Macadam's lab to repeat this work, the results of which are discussed in the next chapter.

4.6 Discussion

An experimental system involving large-scale infection and PCR amplification was optimised and successfully used to amplify and characterise poliovirus intertypic recombinants. Different recombinant populations were amplified from two different regions in the genome. The first region extended from the end of the *VP1*-coding sequence to the CRE, thereby including the last part of the P1 region and the first part of the P2 region (Figure 4-3 A). The second region was located downstream of the CRE, encompassing the last part of the P2 region and the first part of P3 region (Figure 4-3 B). The two possible reciprocal occurrences of recombination were studied from both regions (recP2-R_MD_L, recP2-R_LD_M, recP3-R_LD_M, and, recP3-R_MD_L).

Different recombinants of different types were isolated from both regions. Within the recP2-R_LD_M, 6 different unique precise recombinants were detected. Moreover, 4 different unique imprecise recombinants were found. Three of these were imprecise-deletion recombinants and one was imprecise-insertion. The deleted fragments, among the imprecise-deletion molecules, were found to be varied in their lengths between ~20 – 1100 nts. Interestingly, the imprecise-deletion recombinant with ~1100 nts deletion was found to be in-frame while the remaining 3 imprecise recombinants were all out-of-frame. No correlation, however, was noticed between the location (either on the acceptor or the donor), the type of recombinants (precise or imprecise), and the maintenance of reading frame (in-frame or out-of-frame). This was reflected by the absence of clustering of any of these features in one or near

one region in the genome, rather recombinants seemed to be distributed randomly, though it is acknowledged that the sample size was too small to conduct any statistical analysis of this. Furthermore, the ratio of imprecise recombinants was found to be 1 molecule in-frame to 4 out-of-frame, which when compared to the random expected ratio (1:3) based on the multiple of 3 codon, did not show a significant difference (Chi-Square P = 0.1). These findings suggested that recombination is a random event at this stage of infection. In contrast, the same recP2-R_LD_M was studied by Tolskaya E.A.*et al.*, who concluded that the distribution is nonrandom within this region of the genome (Tolskaya et al., 1987). They suggested that there is a selection against the coding sequence of polypeptide 2A and in favour of the *2B*-coding region. In fact, looking at Figure 4-12, there seems to be more recombinants within the *2B*-coding region and less within the sequence of the polypeptide 2A. Nonetheless, as only few a recombinants were analyzed, this was further analyzed in the next chapter where several hundreds of recombinants were obtained.

Similarly, imprecise and precise recombinants were also successfully analysed from the recP2-R_MD_L (the reciprocal population). Less recombinants were isolated due to the time limitation, from those, 5 recombinants were found to be precise, and one imprecise-deletion recombinant with 20 nts deletion (Figure 4-13). No special clustering was observed; rather recombinants were detected within all the coding regions except the *VP1*-coding region. The lack of recombination occurrences within the *VP1*-coding sequence was also observed in the recP2-R_LD_M (Figure 4-12). This region is part of the capsid region where recombination had rarely been observed to occur (Cao et al., 1993). This stems from the fact that the capsid sequences are the most divergent region between the two viruses used in this study, and folding of capsid polypeptides are extremely sensitive to the change of amino acid sequence (Murray et al., 1988a). At this stage, it was thought that this underrepresentation at the *VP1*-coding region was perhaps due to the small number of analysed recombinants. Nonetheless, this appeared not to be case when the analysis involved a greater number of recombinants (see the next chapter).

Intriguingly, the PCR products of recombinants that have their 5' part derived from PV1 (recP2-R_MD_L) were found to be higher than when the 5' part derived from PV3

(recP2-R_LD_M). This can be seen by the more intense band, which resulted from the former in comparison to the latter (Figure 4-16). This could be explained by the copy-choice mechanism, which is widely regarded as being operative in poliovirus (Kirkegaard and Baltimore, 1986, Lowry et al., 2014). If the cells were co-infected with wild-type viruses that replicated at the same rate, the probability of switching in either reciprocal sense should also be the same. However, the PV3 virus used in this experiment was found to be ~1 log slower than the PV1 in the growth kinetics curve (Figure 4-8), which made the PV1 viruses more available to act as acceptors. Therefore the frequency (reflected by the thickness and intensity of the band on 1% agarose gel) of recP2-R_LD_M was found to be lower than recP2-R_MD_L. The results are consistent with findings by Jarvis and Kirkegaard, who suggested a correlation between recombination events and the replication activity of the recipient genome (Jarvis and Kirkegaard, 1992). Moreover, this was also observed among recP3-R_LD_M and recP3-R_MD_L (Figure 4-14 and Figure 4-15). Nonetheless, it was not as clear as the P2 region, which could be down to the presence of the nonspecific products. Unfortunately, because of the time restrictions, this was not investigated any further.

On the other hand, precise recombinants were found to be predominant within the analysed populations, *i.e.* they occurred at much higher frequency than the imprecise. The ratio of precise to imprecise within recP2-R_LD_M was 11/4 while it was 6/1 within recP2-R_MD_L and 3/0 in the recP3-R_LD_M. The apparent preponderance of precise recombinants was reported in several previous studies (Romanova et al., 1986, Kirkegaard and Baltimore, 1986, King, 1988). The explanation of such an observation may be the selective advantage of the precise recombinants, based on the fact that they possess the same genome length of the parental viruses, hence they may replicate faster than the imprecise. Alternatively, it could be speculated that the high number of precise recombinants may have resulted from a resolution process of early imprecise recombinants (see section 6.2.3).

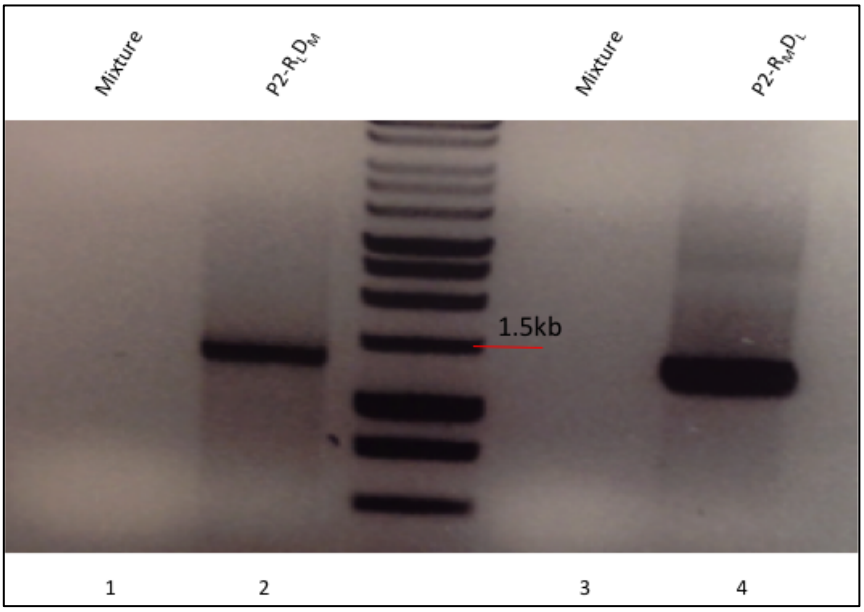


Figure 4-16 Comparison between the two possible reciprocal populations within the P2 region (recP2-R_MD_L & recP2-R_LD_M).
PCR products amplified with primP2-R_LD_M (lane 1, 2) and primP2-R_MD_L, screened on 1% agarose gel, and stained with ethidium bromide. Mix = cDNA from RNA mixture of both viruses

5 Analysis of recombinants by NGS

5.1 Introduction

Chapter four demonstrated and described a new approach for studying recombinants in wild-type poliovirus in the absence of any external pressure. The results gained from cloning and sequencing revealed 18 unique junction sites within an area of 1.3kb (nucleotides 3235-4548), of those, one imprecise-insertion out-of-frame molecule, three imprecise-deletion out-of-frame molecules, and one imprecise deletion in-frame recombinants were found. These findings imply that there were probably many potential recombination sites within the amplified products. It also suggests that 5 hours post-infection – before completion of a full replication cycle – there was already some selection on the virus population.

Recombinants within the targeted region were PCR amplified and NGS technology was used to sequence the amplicons. The amplicons were isolated and sequenced by Illumina MiSeq. Subsequently, the generated reads were analysed by the optimised bioinformatics system discussed in Chapter 3. Because of the time limitation, only the analysis of recP2-R_LD_M (recombination product in which the PV1 (Mahoney) is the donor genome and the PV3 (Leon) is the recipient genome) is demonstrated in this chapter.

This chapter demonstrates the analysis of several hundreds of recombinants analysed 5 hours post infection between two serotypes of poliovirus. The analysis was developed to examine:

a) The randomness.
b) The influence of the secondary structure of RNA on junction position
c) Similarities and sequence identity between both viruses and how they can influence the recombination process.
d) The primary sequence compositions around the junction site, e.g. GC-content, homopolymer, mononucleotides and dinucleotides.

Aims

After succeeding in identifying several different recombinants as outlined in chapter four, the aim was to detect more recombinants by using NGS technology to generate an extensive population of recombination. It was planned to use the optimized bioinformatics described in chapter three to identify the recombinant, coupled with some other bioinformatics software and custom Perl code to characterize the identified recombinants. By performing this detailed analysis, the purpose was to understand the nature of the recombination process at an early stage of the virus cycle. Moreover, such explanatory data would help to differentiate between precise and imprecise recombination and whether they were derived by different mechanisms. The ultimate goal was to determine the factors that could influence a recombination population after five hours of co-infection before the virus was assembled or packaged.

5.2 Generating recombinants samples for the NGS sequencing

As mentioned in chapter 4, GATC Ltd. failed to sequence the samples that were generated at the University of Warwick. Moreover, their preliminary analysis resulted in the destruction of all samples meaning that the experiment had to be repeated. Therefore my supervisor Professor David Evans had arranged a visit for me to Dr. Andrew Macadam's laboratory at National Institute for Biological Standards and Control (NIBSC) institution to repeat the experiment. The same experimental conditions were used at NIBSC (Figure 4-2)

The primers (primP2-R_LD_M) that specific to amplify recP2-R_LD_M were used to amplify recombinants and a band of the expected size was observed (1.3kb) in addition to smears and two secondary bands that presumably corresponded to imprecise recombinants (Figure 5-1). The band below the major one corresponded to imprecise-deletion and the one above the major band reflected the imprecise-insertion recombinants (as found in the data analysis – see below). No band was observed for the control sample wherein an equimolar mixture of parental viral RNA was used as a template for the cDNA synthesis and PCR reactions (see section 4.4.1). The PCR products of both the co-infection sample and the control were then purified to be sequenced by Illumina MiSeq NGS.

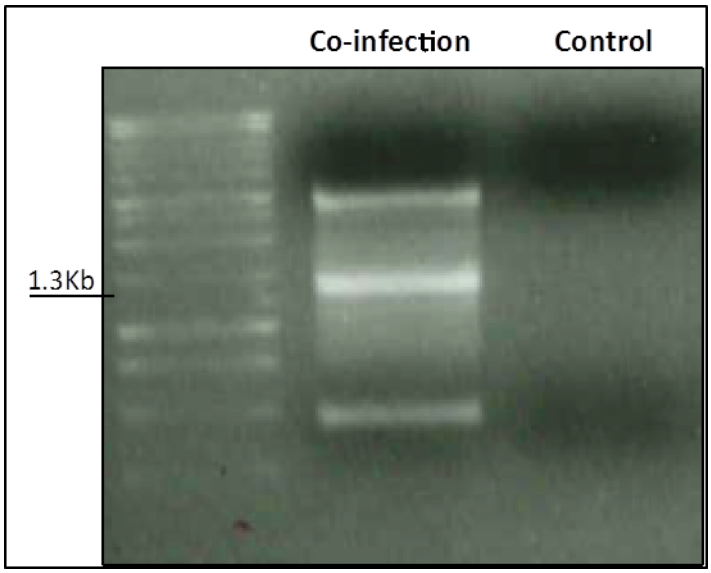


Figure 5-1 Amplicons of precise and imprecise recombinants within recP2-R_LD_M (P2 reagion) generated for NGS sequencing
Viral RNA was extracted 5h post infection, reverse transcribed, cDNA synthesised, and PCR amplified using primP2-R_LD_M. The resulting products were analysed by 1% agarose gel electrophoresis after staining with ethidium bromide. The major band of expected size (~1.3kb) reflects precise recombinants, while the smears and secondary bands represent imprecise recombinants (as found in the analysis). The control lane reflects the absence of artefact wherein an equimolar viral RNA mixture was used as a template for PCR.

5.3 Description of NGS dataset and the analysis methods

NGS data

Library preparation for Illumina MiSeq was performed according to the manufacturer's instructions. Briefly, 1ng was used from the amplicons, simultaneously fragmented and tagged with unique adapter sequences, followed by 10 cycles of PCR, before being loaded into the MiSeq instrument. The generated reads were uploaded to Illumina BaseSpace, a cloud-based genomics analysis and storage platform that directly integrates with all Illumina sequencers.

Duplicates

Duplicates are NGS reads with identical sequence; these reads can be generated during the library preparation, specifically in the PCR step after adding the adaptor sequences. Counting these reads would inflate the proportion of the NGS reads for some recombinants in the dataset. Therefore, they were removed from the recombination dataset, so that we increase the chance that every recombination read observed in the dataset is in fact driven from an independent recombination event.

Flattened vs. Unflattened

The data was divided into flattened and unflattened groups. The former took into consideration the number of unique recombination junctions regardless of their NGS-read representation and the latter accounted for the number of NGS reads. The flattened data could be used to study the presence of any favoured or disfavoured sites for the recombination mechanism across the genome. On the other hand, considering the unflattened data would help in studying the replicative nature of recombinants and/or the bias towards generating a recombination junction independently from the same site. Using the NGS reads to reflect the fitness of the recombinants could be misleading in the presence of duplicates. However, removing the duplicates would render the dataset unable to inform about the replication competent recombinants, since what will remain after excluding the duplicates are unique independent reads and not necessarily derived from the same recombination event.

Differentiating between the technical and biological duplicates would need further verification steps as discussed later in this chapter. Therefore, the analysis described in this chapter informs about either preferred sites (flattened data) and or preferred recombination junctions, that can be independently generated several times (unflattened data).

Junction definition

The junction location reflects the point at which the polymerase switched from donor (PV1) to recipient (PV3) template. In most cases, the junction cannot be unambiguously determined due to local sequence identity between the parental genomes. Therefore, the position of each junction site was assigned to either the 3’ or the 5’ end of the identical sequence. Because no significant difference (in preliminary analysis) was observed between the two locations, the 5’ end of the sequence identity was considered as the junction location and described in this chapter.

5.4 Filtering the reads

Using the BaseSpace software analysis, FastQC and FASTQ ToolKit, the low-quality reads and the reads less than 150 nts were removed from the dataset. Additionally, the regions of low-quality sequence were trimmed from both ends of the reads if they existed. The numbers of remaining reads were 792,437 and 1,484,341 for the control and the sample respectively. The length of the majority of remaining reads was 250 nts (Figure 5-2).

The next step was to filter out the parental reads, which can impose a source of noise when analysing the recombination reads. These exist in the dataset as they were derived from a fragmented 1.3 kb PCR product, the majority of which would generate reads identical throughout their length to the donor or recipient molecules. Using Bowtie aligner (Langmead and Salzberg, 2012), the reads from both the control and the sample were first aligned against PV1, the unmapped reads were then re-aligned to PV3, and the remaining unmapped were kept for recombination analysis (Illustration of the filtration process is in Figure 5-3). A further step to ensure the uniqueness of the NGS reads was performed by removing the duplicates from the recombination reads found in the co-infection sample. Recombination reads left in the dataset constitute 1.21% of total reads and almost zero for the control sample (Figure 5-4). These findings are indicators of the specificity of the PCR approach that was used in this project.

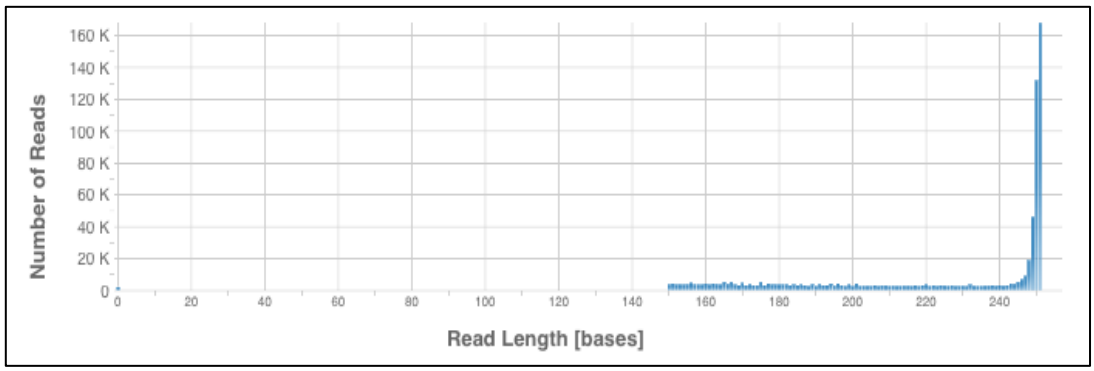


Figure 5-2 Read length distribution after removing short and unqualified reads
The short and unqualified reads were filtered out from the total NGS reads generated by Illumina MiSeq. The remaining reads were binned based on their number (y-axis) and their length (x-axis) The figure was generated by Illumina BaseSpace built-in software (https://basespace.illumina.com/home/index).

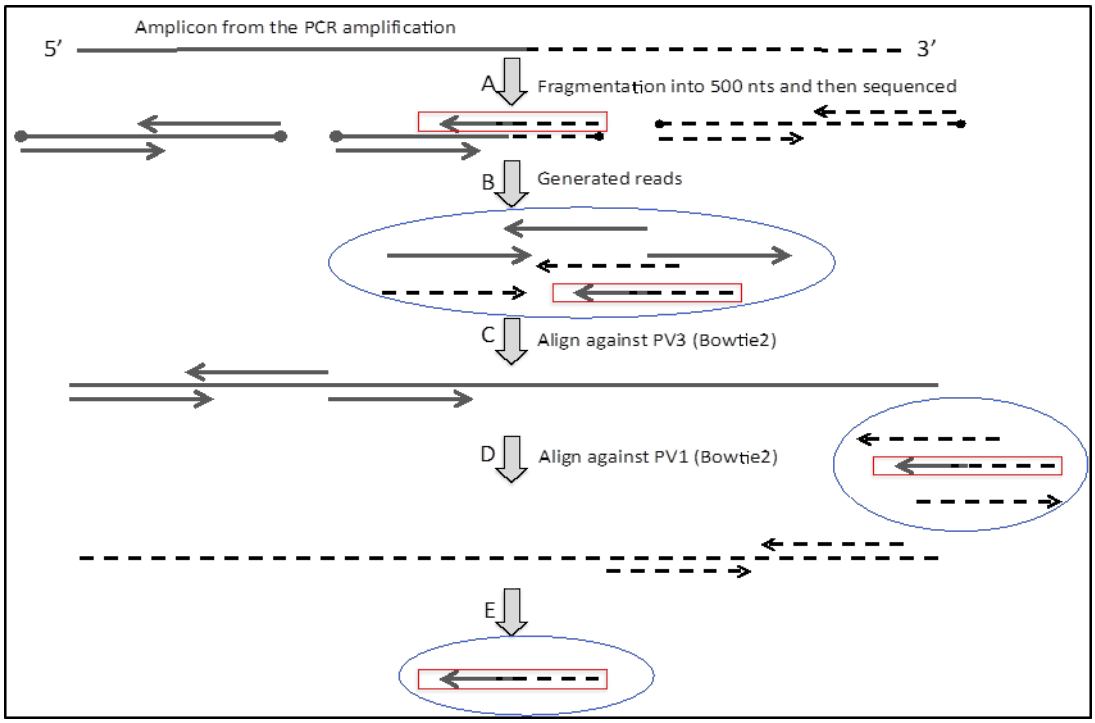


Figure 5-3 The pipeline of removing parental NGS reads from the dataset
An illustrative drawing explaining the steps of generating and removing the parental reads from the dataset. A two-halves line depicts the recombinant amplicon; the solid-grey represents the acceptor while the dashed-black represents the donor. The horizontal arrows refer to the NGS reads whereas the thick vertical arrows refer to the steps of the pipeline. The lines ending with filled circles denote the 500 nts fragments. The PV3 acceptor genome is represented by a solid grey complete line, and the dashed black complete line represents the PV1 donor strand A) The recombinant amplicon is fragmented into 500 nts during the library preparation. B) The majority of resulted reads are derived from parental genomes. C) Filter out the reads that derived from the acceptor strand by aligning all the reads against the PV3 strand, and keep the unmapped reads for the next step. D) Filter out the reads that derived from the donor strand by aligning all the reads against the PV1 strand, and keep the unmapped reads. E) What is left are the reads derived from recombinants.

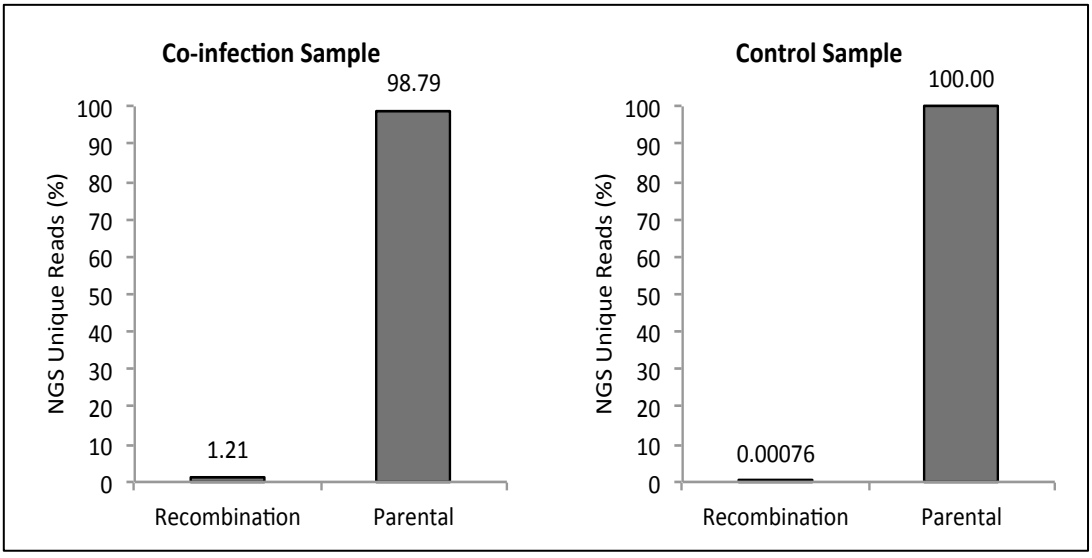


Figure 5-4 Recombination NGS reads percentage in comparison to parental genome read
The grey bars represent the number of reads, and the numbers above them refer to the percentage of the reads. The left panel corresponds to the coinfection sample, while the right panel corresponds to the control no-coinfection sample.

5.5 Finding Recombinants

The 1.12% recombination reads were analysed by ViReMa algorithm at mode N0 X5 as described in Chapter 3. Subsequently, the recombination reads were manually securitized by extracting the reads represented at each junction and searching for recombinants by either BLAST or Bowtie2. As for BLAST, reads from several recombination junctions were extracted, and used in BLAST search via the NCBI website. The reads would be accepted as reads derived from a recombinant molecule if the resulting alignment showed a clear pattern of mismatches when aligned to one of the parental genome.

On the other hand, reads represent several recombination junctions were aligned by Bowtie2 against one of the parental genomes using Bowtie2 with configurations that increased the mismatch tolerance (see section 2.8.6). Because the recombinant molecule genome has two parts, each derived from either parent, the part that does not belong to the reference parental genome in use will demonstrate a pattern of mismatch when visualised in a genome viewer (Figure 5-5). The recombination junction was accepted if a clear pattern of mismatch was observed.

Furthermore, to be included in the subsequent analysis, the observed recombination event should have forward and reverse reads if it lies towards the middle of the amplicon, and either forward or reverse reads if it is located towards the ends of the amplicons as it was verified by the bioinformatics system. Finally, the recombination events that were represented by less than 3 reads were excluded from the dataset. This to increase the confidence that the recombination event was not mistakenly reported based on mismatches (the cut-off number was arbitrarily chosen).

The number of the unique recombination events that passed all the criteria is 313 represented by 16860 unique reads; two-thirds of the recombinants were found to be precise recombinants, with the rest imprecise (Figure 5-6). Among the imprecise recombination junctions, the imprecise-deletion recombinants were found to be 2.4 fold higher than the imprecise-insertion recombinants (Figure 5-6). It is clear that the precise recombination is occurring at a higher frequency than the imprecise and the imprecise-deletion overshadows the imprecise-insertion. Whether these variations

are influenced by biological factors will be further investigated in the following sections.

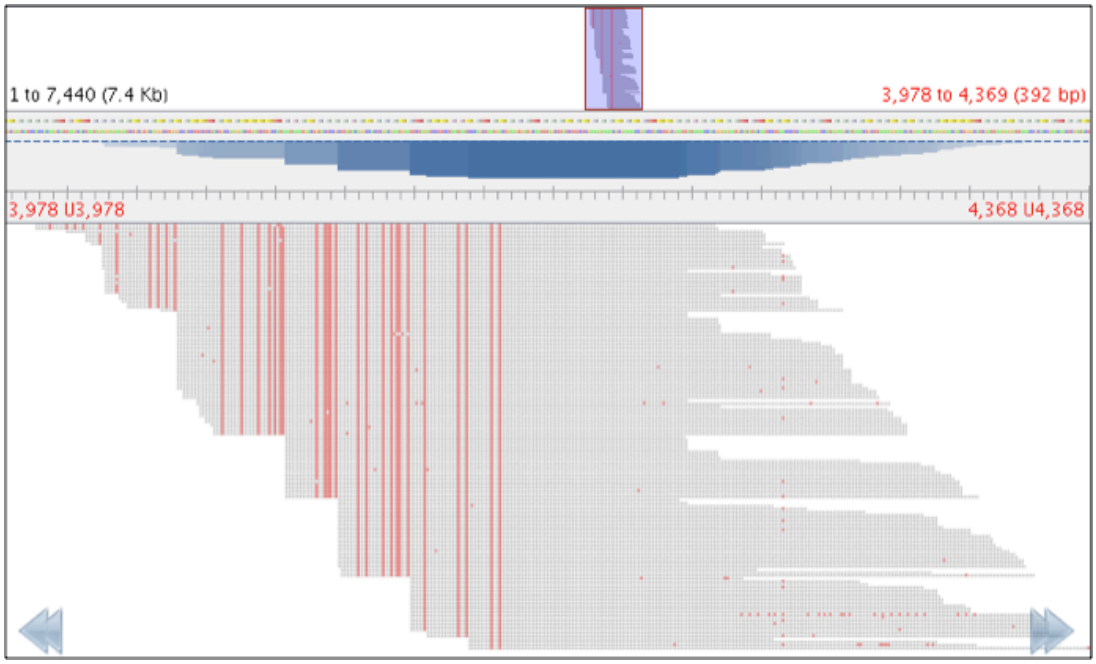


Figure 5-5 Bowtie local alignment of recombinant's reads against PV1 parent
Alignment carried out by Bowtie2 using settings that allow mapping in the presence of mismatches, then visualised by Tablet (see section 2.8.13). The figure represents a screen-shot for an example recombination alignment generated by Tablet. The vertical red lines represent mismatches, which correspond to PV3 parent, and the grey part is the PV1.

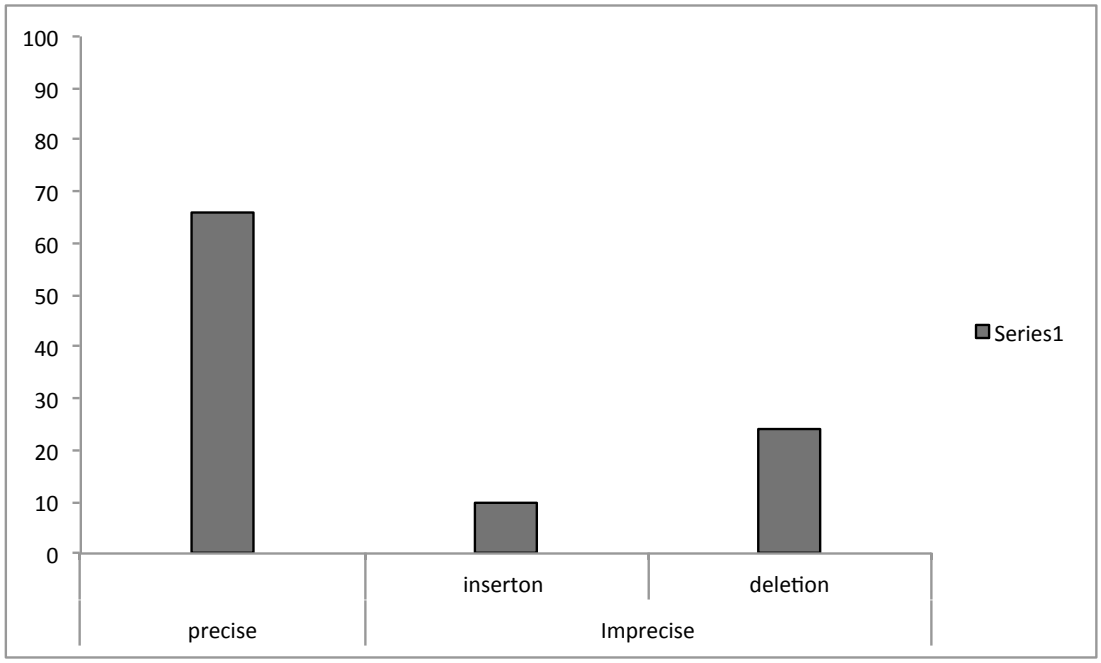


Figure 5-6 Precice and imprecise NGS reads percentages
The grey bars represent the percentage of unique recombinants found in the dataset for each type of recombination. The X-axis defines the type of recombination tested and the y-axis demonstrates the percentage of the unique junction in the total 313 recombination junctions found in the database.

5.6 Visualising recombinants found by NGS on the Parallel Coordinates visualisation tool

To screen the locations and the features of the 313 identified recombinants, the Parallel Coordinates visualisation tool (see section 2.8.14) was further developed and manipulated to fit all the recombinants generated from the NGS sequencing. In this visualisation method, recombinants are represented by lines that connect the switching location of RdRp on the donor strand with its corresponding landing location on the acceptor strand.

This visualisation tool allows for tracing recombinants between the two genomes and links each recombinant with its related features such as the number of NGS reads, being in-frame or out-frame, and the length of the inserted and deleted fragment among the imprecise recombinants. Using Parallel Coordinates visualisation tool, three different recombination maps were established, precise, imprecise-deletion, and imprecise-insertion (Figure 5-7).

Precise Recombination Map

In this map all the lines are horizontal (no slope) and because the blue colour was assigned for the in-frame recombinants, all the lines in this map appeared in blue colour; this because all the precise recombinants by definition are in-frame. Purple colour was assigned for the recombinants that were found in the laboratory at the University of Warwick via the cloning and Sanger sequencing (see section 4.4.3), and was found again among the NGS sequencing reads that were generated at NIBSC. This finding verifies the reproducibility of the method used in this research. The number of NGS reads that represent each recombinant is demonstrated on the last vertical bar on the right, which is termed ‘frequency’. The vast majority of the recombinants are represented by 3-300 NGS reads while fewer were found between 300-500 reads and one recombinant was observed at a frequency of 823 reads (see 5.7.2.)

In a previous study targeting the same region of the genome, it was suggested that intertypic precise recombination was nonrandomly distributed (Murray et al., 1988b,

Freistadt et al., 2007b). Moreover, a study by Lowry *et al* , in which the same region was analyse, showed distinct clustering for recombinants (Lowry et al., 2014). However, apart from the *VP1*-coding region, identified precise recombinants in this project appeared to occur everywhere across the whole targeted region without any apparent clustering.

This discrepancy may be due to the differences in methods used to isolate recombinants; they selected for viable recombinants - that had time to be packaged - with mutant markers. In contrast, the virus progeny was harvested early in the infection and no selection of any kind was used in this project and no viability was required.

Imprecise-deletion Map

Imprecise-deletion, by definition, can be either in-frame or out-of-frame. In this map, many red lines can be observed representing the out-of-frame imprecise-deletion recombinants (Figure 5-7, B). The recombination lines between the parental genome are not horizontal anymore; they became diagonal, moving downward from the donor to the acceptor genome reflecting the path of RdRp. In imprecise-deletion recombinants the RdRp should move towards the 5' end of the acceptor genome to skip part of the genome, the slope of the line reflects the length of the deleted fragment *i.e.* the greater the slope, the larger the deleted fragment. Some lines look like horizontal lines because of the scale of chart, but they are not. In fact, these lines represent recombinants with small deleted fragments between 1-10 nts. The third vertical bar from the left, which is called 'length', represents the length of the deleted fragment, over which a clustering at a short-length deleted fragment can be observed. Finally, as before, the last vertical bar on the right of the chart represents the number of NGS reads that represent each recombinant. It is worth noting that imprecise-deletion recombinants are occurring cross the whole targeted genome apart from a ~350 nts region in the 2A coding region, which seems to be a "cold spot" (see section 5.8 & 5.12)

Imprecise deletion recombinants are historically referred to as defective interference particles (DI) in the literature (see section 1.3.6). These particles have deletion located in the capsid-encoding region (Cole and Baltimore, 1973b) and still they can replicate. However, because they don't synthesise virus proteins they cannot produce virions, and rather compete for the wild-type (helper virus) structural protein (Cole et al., 1971). DIs are proposed to be an imprecise recombination mediated either by the copy-choice mechanism (Lundquist et al., 1979), or the loop-out mechanism (Kuge et al., 1986). In all cases, DI genomes have always been studied as a result of intratypic recombination, and have never been reported in the P2 or P3 region. The results demonstrated in this chapter are strong evidence that deletion can happen in the middle of the polyprotein region (P2), and more importantly that it can result from an intertypic recombination.

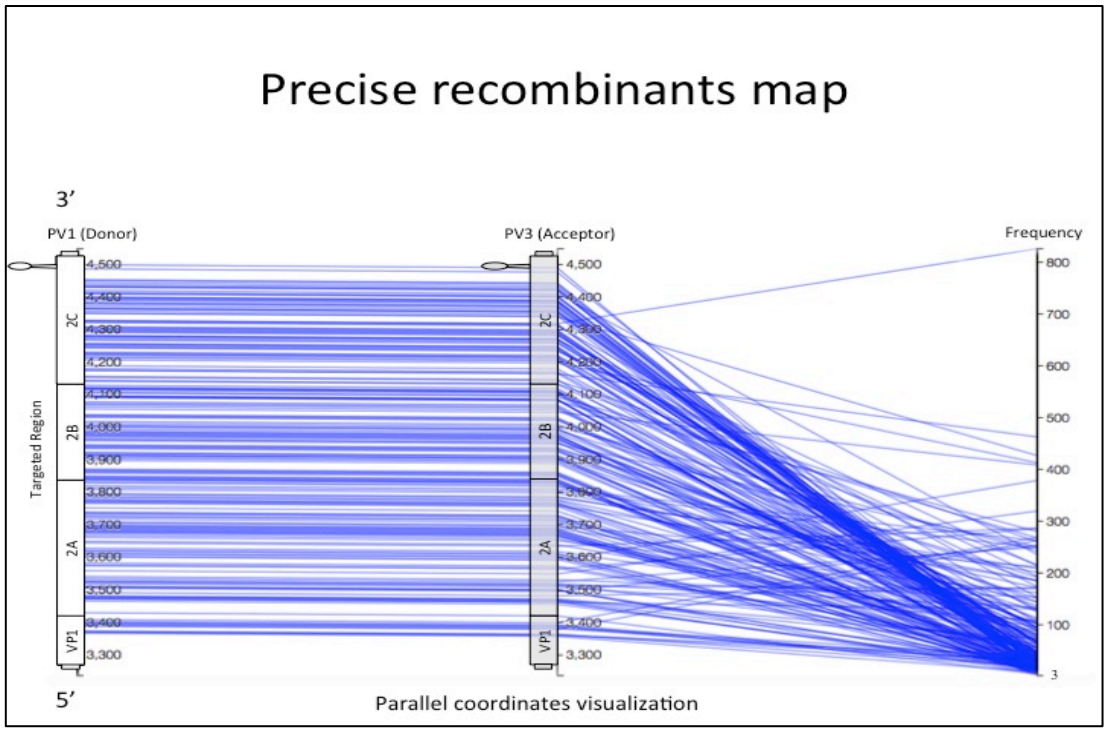
Imprecise-insertion recombination map

The lines are moving diagonally in the opposite direction to the imprecise-deletion map, this because the RdRp in this situation duplicates part of the genome, thus it switches towards the 3' end of the acceptor (Figure 5-7, C). Three recombinants were detected have their acceptor strand locations after the CRE, which is outside of the PCR range of this project. Although, the RdRp switched templates outside the PCR-region range in these three molecules, the primer sites remained preserved. This is not the case with imprecise-deletions where switching templates outside the range would result in losing the 3' primer site. Interestingly, the first half of 2A (~220 nts) region in the donor strand seems to be selected against, which also was observed in the imprecise-deletion (Figure 5-7, B). The lack of recombination at this region would imply a particular feature(s) retained by this region that would affect the formation of recombination. The fact that this was not found among the precise recombinants (Figure 5-7. A) suggests that it is specific for imprecise recombination.

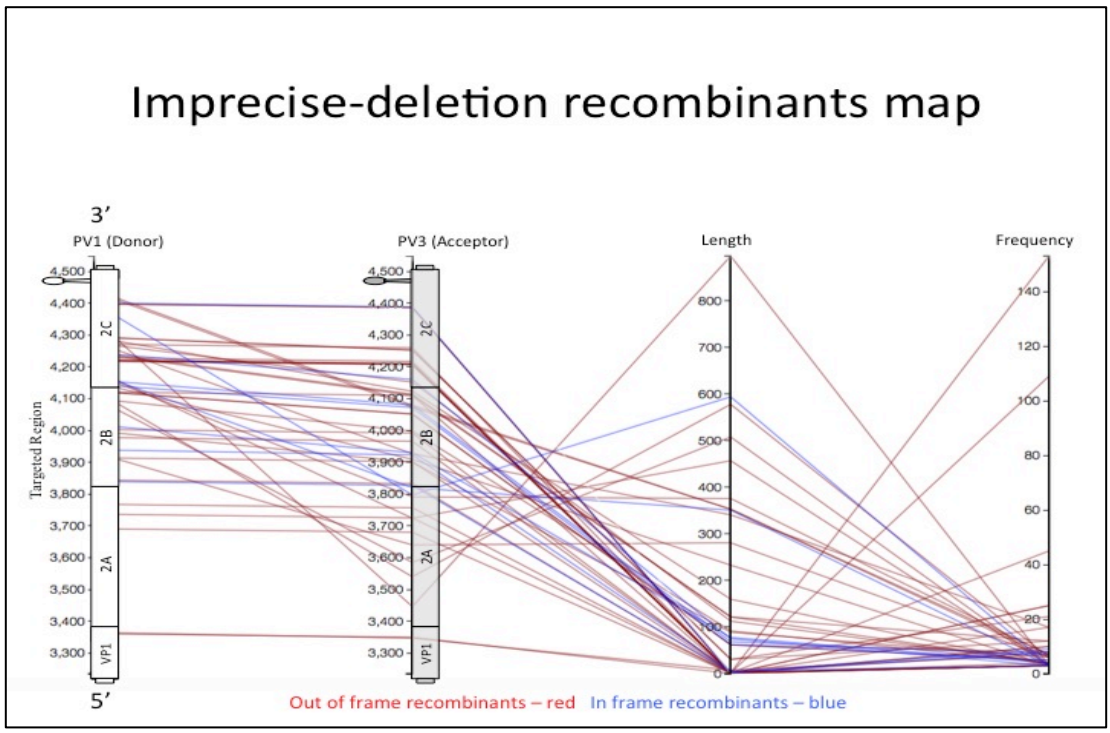
Imprecise-insertion recombinants are rarely reported in the literature, either in natural isolates or cell culture. The difficulty of isolating recombinants stems from the fact that they occur at a low frequency. Together with being less fit than the precise recombinants, the imprecise-insertion recombinants - as they possess longer genomes than the wild type - disappear after several hours (within the host) or serial passage due to the inevitable competition with the parental viruses and the precise

recombinant viruses. However, Lowry et. al were able to isolate viable imprecise-insertion recombinants at the same region of the genome (P2). The insert was deleted upon serial passage and restored the original wild-type genome length. This occurrence was also observed in an engineered virus carrying an insert (Alexander et al., 1994). The results in Figure 5-7, C show that imprecise-recombination does in fact occur at an early stage of infection in a random manner, regardless of their viability.

A)


Precise recombinants map
3′
PV1 (Donor)
PV3 (Acceptor)
Frequency
Targeted Region
2C
2B
2A
VP1
5′
Parallel coordinates visualization

B)


Imprecise-deletion recombinants map
3′
PV1 (Donor)
PV3 (Acceptor)
Length
Frequency
Targeted Region
2C
2B
2A
VP1
5′
Out of frame recombinants – red
In frame recombinants – blue

C)

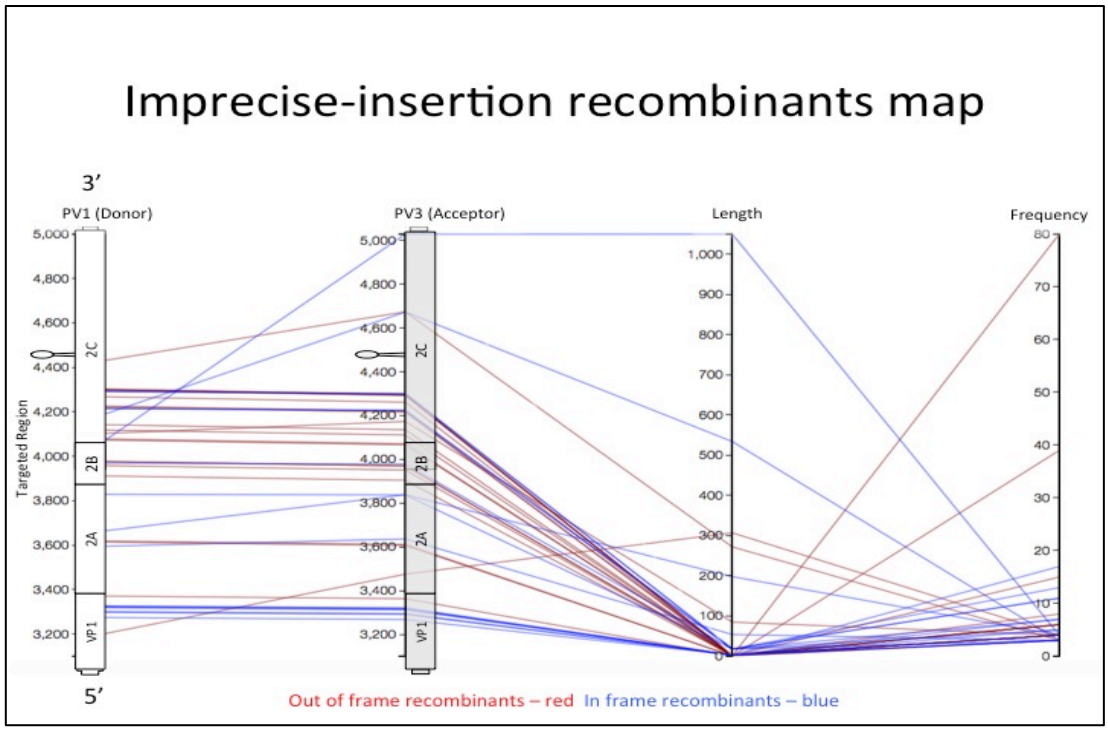


Figure 5-7 Precise and Imprecise recombination maps (Parallel Coordinates Visualisation)
The identified recombinants and their corresponding features were plotted between the two parental genomes as coloured lines in a parallel coordinates style. Each line represents an individual recombinant. The targeted genome of the PV1 (donor) strand is represented by a vertical white cartoon drawing with protein coding region indicated while the targeted genome of the PV3 (Acceptor) strand is represented by a vertical transparent-grey cartoon drawing. The locations of the lines on the PV1 strand are the positions where the RdRp has switched template whereas the locations on the PV3 strand are where the RdRp has re-associated. A name was assigned for each vertical bar; 'frequency' = the number of unique NGS reads, 'length'= the inserted or deleted fragment length. The direction of the lines indicates the path of the RdRp. In the precise recombinants, the lines are horizontal with no slope as no deletion or insertion are involved. In the imprecise-deletion, the lines are diagonal pointing downward as the RdRp will jump towards the 5' end on the acceptor whilst in the imprecise-insertion the lines are pointing up indicating the move of RdRp that will create an insertion in the resulted recombinant molecule. As the effect of the scaling view, the short insertion or deletion lines appear as horizontal as the precise ones. The red lines refer to out-of-frame recombinants and the blue lines indicate the in-frame recombinants. A) Precise recombination map. The purple lines are recombinants identified previously by cloning and Sanger sequencing and were found again by NGS. B) Imprecise-Deletion map C) Imprecise-Insertion map.

5.7 Test of randomness of recombination

In this test the overall distribution of recombination junctions was studied to examine if what was observed in the maps were random events. The findings were divided into three samples: precise, imprecise-deletion and imprecise-insertion, and each of these were analysed as both flattened and unflattened data, then different statistical methods were used to compare the distribution with a random model (see section 2.6). The tests in this section studied the randomness of the recombination based on their locations on both viruses, the donor and the acceptor. However, as no significant difference was found between both locations in terms of randomness, the locations on PV1 was described in this section as a representative aspect for the randomness.

5.7.1 Precise recombination randomness (flattened data)

A random model was established by simulating random precise recombination reads with the same length (250 nts) in the same-targeted region (this randomisation was done by Excel), and ViReMa was used to find the junctions of the simulated reads. This was treated as a random-control model against which the sample was compared. This test ignores the number of NGS reads for each event, and rather focuses on the recombination event at each location. Starting the counting of recombination events from the beginning of the amplicon where no event was observed (zero), moving towards the 3' end of the targeted region, the first encountered event at the next location would be counted as 1. Moving to the next location with the counter set to 1, which will remain 1 as long as no recombination event is encountered. Once a recombination is met then the counter will become 2. The same counting process continues until reaching the end of the amplicon. The resulting numbers from the counting process - for both the sample and control - were then normalised to 100% and plotted on a chart (Figure 5-8, A).

The null hypothesis (H_0) is that the precise recombination sample and the random-control sample come from the same distribution, and the alternative hypothesis (H_a) is that they come from different distributions. Testing the goodness of fit by Mann-Whitney U test resulted in a significantly small P value ($P<0.01$), which meant that

the chance that both samples come from the same distribution is very low; therefore the null hypothesis was rejected. Looking at the chart in Figure 5-8, A, the distributions of both samples are deviating at the beginning of the amplicon, specifically over the *VP1* region and then moving towards each other until they meet at the end of amplicon.

This significant absence of recombination over the *VP1*-encoding region among the precise recombination could be due to the physical location of the *VP1* being at the 3' end of the amplicon, rather than being selected against for a biological reason; as a result it is underrepresented in the data (see section 3.7.1). To test this, the closest recombination occurrences to both ends were compared using Equation 3-3. The first detected recombinant from the 3' end is at location 4499, which is 49 nts away from the 3' end of the region amplified. Whereas, the first detected recombinants from the 5' end (the location of the *VP1*) is at 3366, which is 131 nts away from the beginning of the amplicon. Using the equations the theoretical sensitivity for ViReMa was found to be 29.6% and 62.4% at location 4499 and 3366 respectively. If the physical location were the reason behind the underrepresentation, then ViReMa should have had detected a sign of recombination at location ~3297 where the sensitivity is equal to 29.6% from the 5' end. The fact that ViReMa needed a 1.1-fold increase in theoretical sensitivity to detect a sign of recombination at the 5' end in comparison to the 3' end suggests that this underrepresentation is due to a biological reason. The selection against the *VP1* region is not surprising, as it is part of capsid-encoding regions, within which recombination has rarely been observed in previous studies (Heath et al., 2006).

The test was repeated after removing the *VP1* region (Figure 5-8, B) and the two samples were found to be derived from the same population; the (H_0) was accepted (P=0.17, Mann Whitney U test). These findings suggest a strong selection against the *VP1* region. Moreover, it demonstrates that the distribution of the precise recombinants occurs randomly cross the P2 region if the *VP1* is excluded.

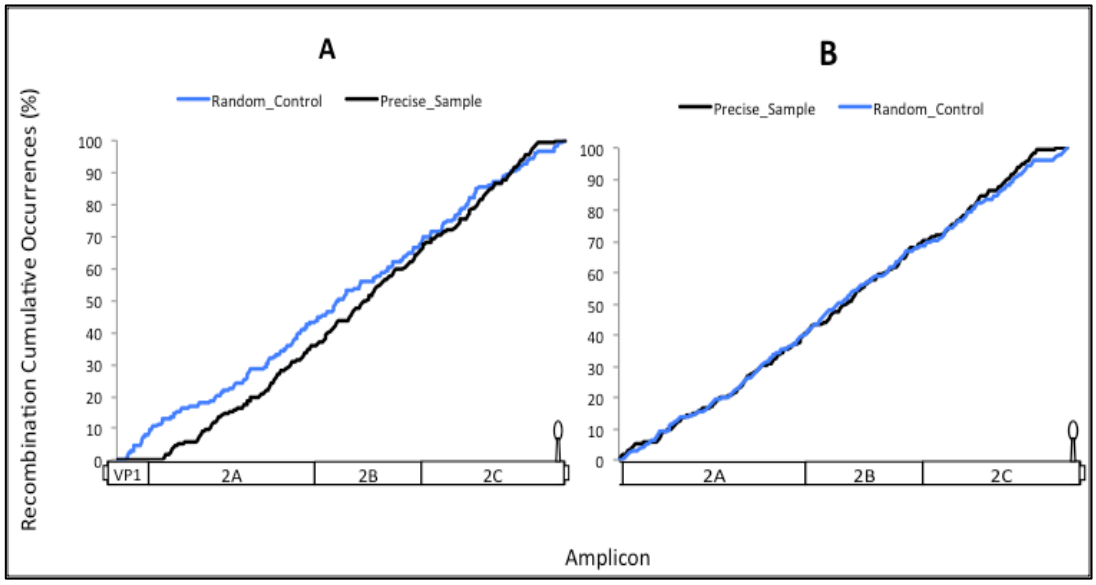


Figure 5-8 Randomness of precise recombination flattened data
The occurrence of recombination was given a score 1 while no occurrence is 0, therefore the number of recombination occurences was calculated in a cumulative way *i.e.* the occurrence of recombination at each postion is added to the occurrence at the previous postion. This was applied on two populations; the precise recombiantion sample and a random recombination model generated by Excel. The findings (y-axis) were plotted against the locations on the PV1 (donor) amplicon (x-axis). The blue line represents the precise recombination random model and the black one for the sample (observed). A) Comparison between both populations over the whole targeted region. B) Comparison between both populations in the absence of *VP1*-encoding region.

5.7.2 Precise recombination (unflattened)

The number of NGS reads - without the duplicates - for each recombinant can be used to test if there is any preference towards a specific recombination junction over the others. However, based on the fact that it is not possible to unambiguously determine the exact location for many recombinants due to the presence of identical sequence stretches between the two viruses (so-called ambiguous junctions), it is important to test how safe it is to consider the number of NGS reads as an indicator for preference occurrences for some recombinants. If two recombination junctions happened to occur over two different places in the middle of an identical sequence, ViReMa will push the findings towards the 3' end of the identical sequence and report them both as one recombination event (Figure 3-18). To test this, a random model was created by simulating random recombination events, taking into consideration the number of NGS reads into account this time (this was performed by the Perl code Appendix 1). Subsequently ViReMa was used to find the recombinants and their corresponding NGS reads. The normalised results were then compared between the random model and the sample (Figure 5-9).

Looking at the results, it is visually clear that the distributions are similar for the random-model and the precise-sample, apart from the *VP1* region and the primers sites at the end. The null hypothesis was rejected ($P<0.001$ Mann Whitney Test) without any modification. While after removing the primer sites and the *VP1*, the null hypothesis can be accepted ($P = 0.08$ Mann Whitney Test).

After examining some of the recombination events, it was found that the number of reads is actually reflective of a systematic bias rather than a biological indicator. For example, the highest number of reads over the *2C* encoding region corresponded to a recombination event that occurred within a 21nt identical sequence, which implies that possibly this is a representation of several events occurring at this region (see section 5.10). Based on these findings it was decided not to consider the number of NGS reads as an indicator of a favoured situation for recombination.

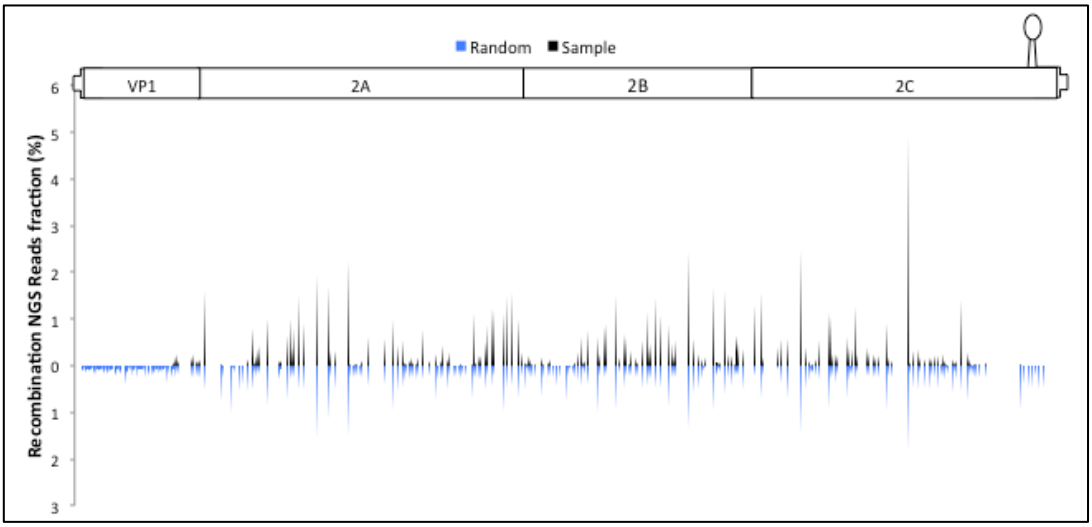


Figure 5-9 Randomness of precise recombination unflattened data
Comparing the number of NGS reads (frequency) for each recombinant found in the precise recombination sample and in the random recombination model**.** Blue refers to the random model and black to the precise recombination sample. The x-axis label is placed on the top of the graph as a cartoon representing the PV1 genome placed in a position that corresponds to the scale. The y-axis is the percentage of the NGS of each recombinant in regards to the total number of reads.

5.7.3 Imprecise Recombinants (flattened)

The imprecise recombination in the case of flattened data was tested based on their ability to maintain the open reading frame. If these molecules occur randomly, then it is expected to see two-thirds of the molecules in each category out-of-frame, based on the fact that the first two nucleotides –within a codon - if disrupted will generate out-of-frame recombinants. The recombinants were divided into two categories, imprecise-deletion and imprecise-insertion, then a goodness of fit was tested by Chi-Square Test between the expected and observed. The results in Figure 5-10 show that the imprecise-insertion is occurring randomly (P=0.88) while, surprisingly, the imprecise-deletion is biased towards the out-of-frame molecules (P=0.0004). The (H_0) is accepted in the former case and rejected in latter case. This is also reflected by the overwhelming presence of the red lines over the blue lines in the imprecise-deletion map (Figure 5-7, B) in comparison to the imprecise-insertion map (Figure 5-7, C). It is not clear if this bias among the imprecise-deletion is due to a selection mechanism, or it is an effect resulting from the sequencing technology or the random sampling. This can be verified by repeating the experiment and comparing the results.

5.7.4 Imprecise Recombinants (unflattened)

In the case of the unflattened data, a random model was established by simulating random imprecise-deletion and insertion reads (by Perl code Appendix 1). The number of simulated reads for both populations is equal to the number of reads found in the sample. ViReMa was used to find the simulated junctions in both cases. The random-control sample and the imprecise sample were then plotted against each other.

The junctions (the location is based on the PV1 genome) from the random-control found by ViReMa were distributed equally throughout the genome in comparison to the imprecise samples (Figure 5-11), which implies the absence of the systematic bias. Based on this, and on the fact that all the duplicates were removed from the dataset, it seems that some other factors contribute to the high number of NGS reads at some locations in the imprecise sample (see next section). A high number of NGS reads was detected at the *VP1* region among the imprecise-insertion, the same region

that was found to be selected against in the precise recombinants. Interestingly, the vast majority of these are in-frame (7/10) with a short insert, which may suggest a preferable site for the imprecise-insertion (Figure 5-7, C).

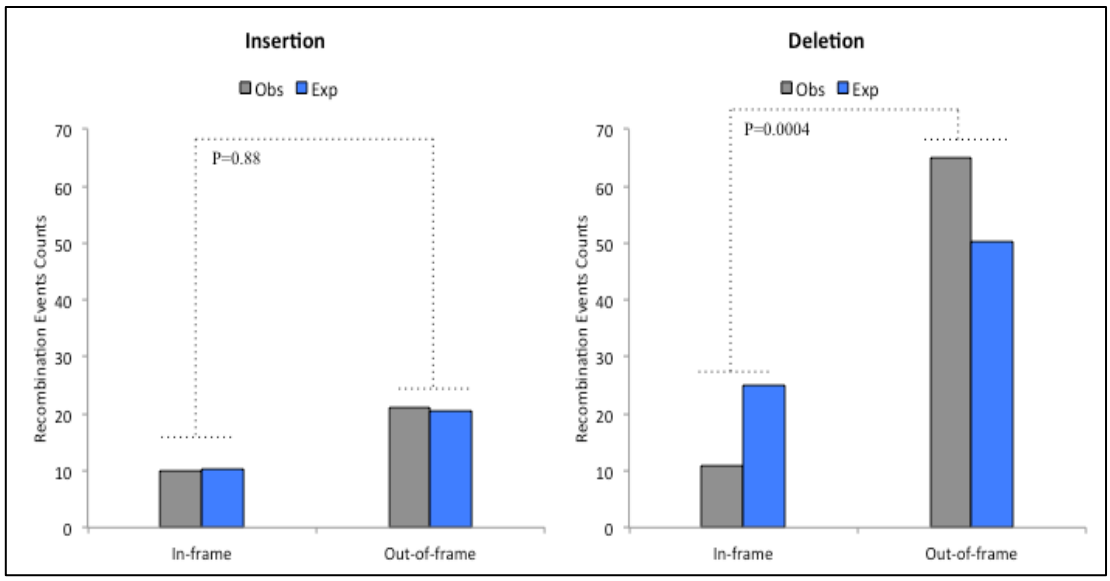


Figure 5-10 Randomness of imprecise-insertion and imprecise-deletion (flattened data)
Blue bars = expected, dark-grey= observed. The dotted line between the bars reflects the difference between the findings with the P value calculated by Chi-Square. X-axis defines the population while y-axis represents the number of recombinants found for each population. The legends and the titles are at the top of each figure. The left panel compares the number of the observed out-of-frame/in-frame recombinants in imprecise-insertion with the expected value, which was calculated based on the whole position (2/3 expected to be out-of-frame). The right panel compares the number of the observed out-of-frame/in-frame recombinants in imprecise-deletion with the expected.

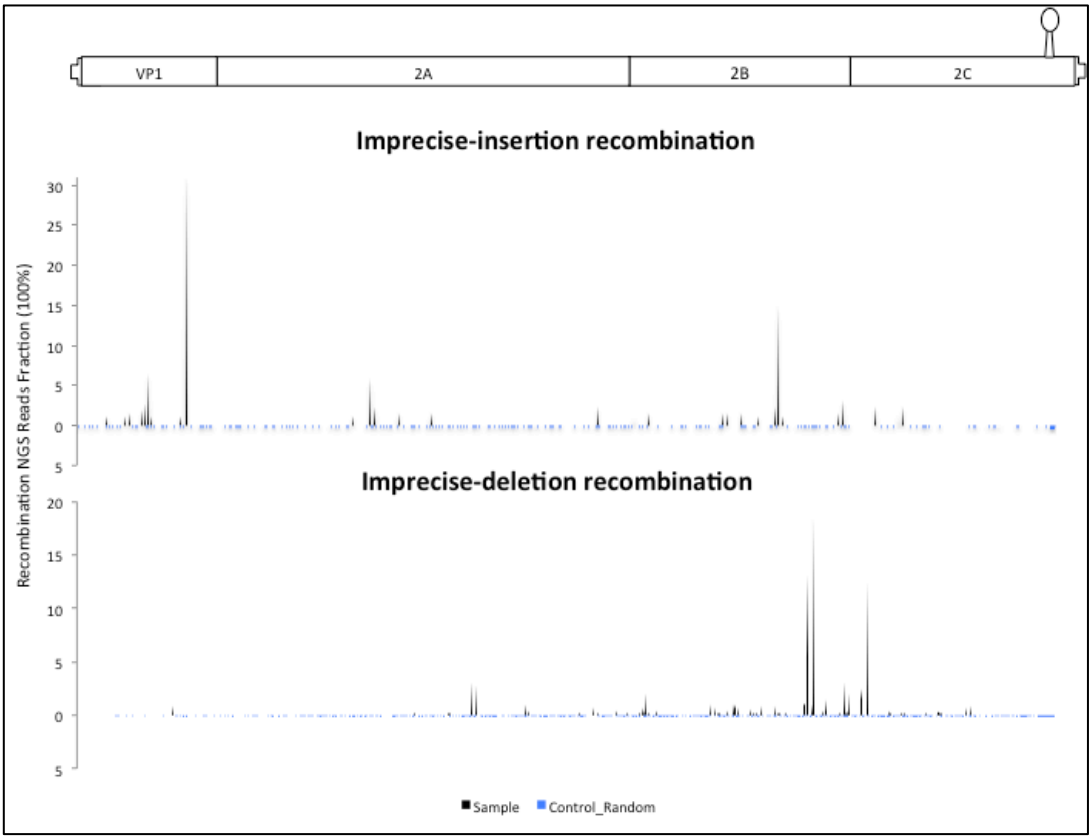


Figure 5-11 Randomness of imprecise-insertion and imprecise-deltion (unflattened data)
Comparing the number of NGS reads (frequency) for each recombinant found in the imprecise recombination (deletion and insertion) sample with the random recombination model**.** Blue refers to the random model and black to the imprecise recombination sample. The graph is divided into two parts sharing the same x-axis label which is placed on the top of the graph as a cartoon representing the PV1 genome placed in a position that corresponds to the scale. The y-axis is the percentage of the NGS of each recombinant in regards to the total number of reads. The sharped bins refere to the recombination events. As there was no clustering among the control samples the bins are appearing small and scattering all over the targeted region.

5.8 RNA Secondary Structure analysis of recombinants found by NGS

Previous studies have implicated local RNA secondary structure as an influencing recombination (Freistadt et al., 2007b, Dedepsidis et al., 2010, Runckel et al., 2013). To further investigate this, the amount of RNA structure on the positive sense genome was defined in the acceptor and donor genomes. This was achieved by calculating the mean folding energy difference (MFED) - a measure of sequence-independent localized RNA structure - between the native sequence and the sequence order randomised control over a sliding 100nts window (stepped by 30 nts) spanning the region within which recombination could occur (this was done by Prof. David Evans). The MFED values - positive indicates sequence order dependent structure, negative values the absence of structure - then were plotted against either the number of recombination junctions within the same sliding window (flattened data) or the number of unique NGS reads (unflattened data). Since the widely accepted model of recombination in RNA viruses suggests that recombination occurs during the negative strand synthesis, the analysis was focused on the structure formed in positive-sense RNA. Recombinants were divided into precise, imprecise-deletion and imprecise-insertion. The Spearman's Rank Correlation test was used to study any correlation with RNA secondary structure for each category individually.

The flattened data of precise recombinants showed a significantly weak negative correlation with the RNA secondary structure in the acceptor strand ($r = - 0.328$, $P = 0.03$) (Figure 5-12, B) while no correlation was found in the donor strand (Figure 5-12, A). From section 5.7.2, the unflattened data for precise recombinants was demonstrated to be a reflection of a systematic bias due to the presence of identical short sequences between the two viruses. Nonetheless, this was also examined for any correlation with the RNA secondary structure, and the results did not show any significant association with either genome (data not shown). The negative correlation found with the acceptor strand may imply that a minimal RNA secondary structure can influence the sites of recombination, and cast more importance on the RNA structure of the acceptor strand than the donor strand within the process of recombination. Nonetheless, as the correlation is weak ($r = - 0.328$), this might be reflective of the analysis method rather than a biological correlation. Therefore, more analysis of further recombinants is needed to verify this.

On the other hand, neither flattened nor unflattened data has shown any correlation when the imprecise-deletion and imprecise-insertion were treated as one sample. When analysed separately, imprecise insertion also did not show any correlation (Figure 5-13) while a weak negative correlation in the donor strand was found in the imprecise-deletion recombinants for both flattened and unflattened data (Figure 5-14, A and C). The negative correlation is attributable to the absence of a recombination junction in the first half of the 2A coding-region, which was found to be a highly structured region; in fact, this region demonstrates the highest MFED value in the targeted genome of PV1 (donor) genome. This is coupled with the notable high occurrence of recombination over the region with low structure within the *2C*-encoding region in both cases.

Some sort of detectable limitation of the assay being used might have caused the gap in the *2A*-encoding region. For imprecise-deletion recombinants to be generated, the RdRp must move towards the 5' end of the acceptor strand, thereby skipping part of the genome. This means if the polymerase happens to switch template at this area, it may re-associate on a region outside of the PCR detectable range used in this project. In addition to this, the correlations are not strong in either the flattened ($r = -0.365$) and the unflattened ($r = -0.388$) data. Therefore it is difficult to prove that the mechanism of imprecise-deletion is influenced by RNA structure in the donor strand, or that the high frequency of imprecise-deletion over *2C*-encoding region is caused by a preferable site defined by the RNA-secondary structure.

At this stage, it appears more convincing that there is no difference between the recombination mechanisms of the imprecise-deletion, and imprecise-insertion, especially that there is no obvious relation between the locations and the occurrences of recombination (Figure 5-7, B and C). The blue and red lines seem to scatter randomly over the genome, apart from the *2A*-encoding region in imprecise-deletion (cold spot) and *VP1*-encoding region in imprecise-insertion (hot spot). However, these exceptions are likely to be a bias caused by the detectable-range effect of the assay, particularly in the case of deletion. Additionally, no special relation between maintaining the open reading frame and the frequency was found, some of the out-of-frame are occurring at a higher frequency than the in-frame molecules, which may

rule out the replication ability as a factor influencing the frequency, which in its turn is more likely to be excluded as the duplicates were removed from the data. Therefore, the imprecise-deletion and insertion were merged into one sample for subsequent analysis, unless mentioned otherwise.

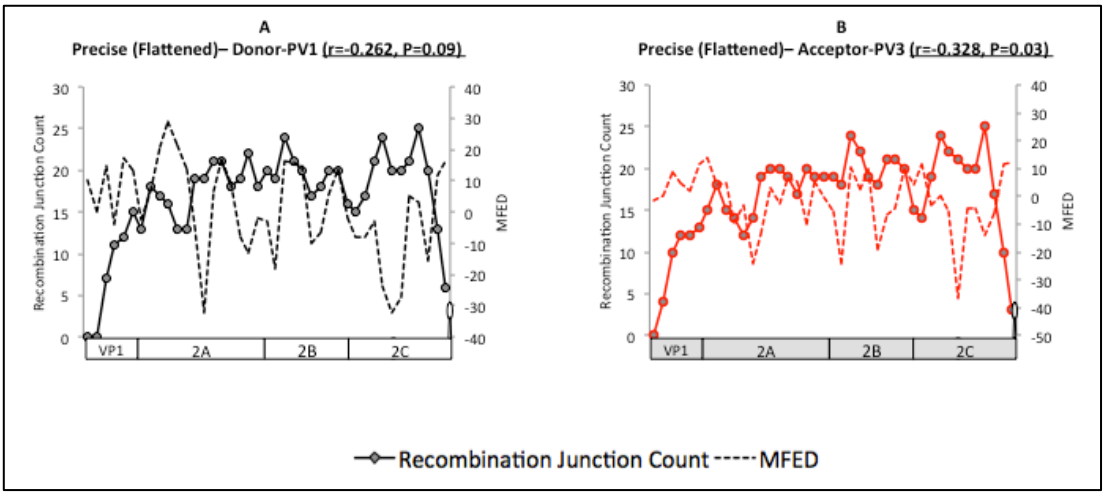


Figure 5-12 Precise Recombination RNA secondary structure analysis (flattened data).
The mean folding energy difference (MFED) was measured within 100 nts sliding window (in 30 nts increments) in the sense strand of parental genomes indicated in dotted line and represented at the secondary y-axis. The solid-marked line represents the recombination junctions count within the same window represented by the primary y-axis. The x-axis was replaced by targeted genome calibrated to the scale. White genome = PV1 (donor), grey genome = PV3 (acceptor). The title above each figure refers to the population's name in its first half, and to the Sperman's Rank Correlation Test results in its second half.

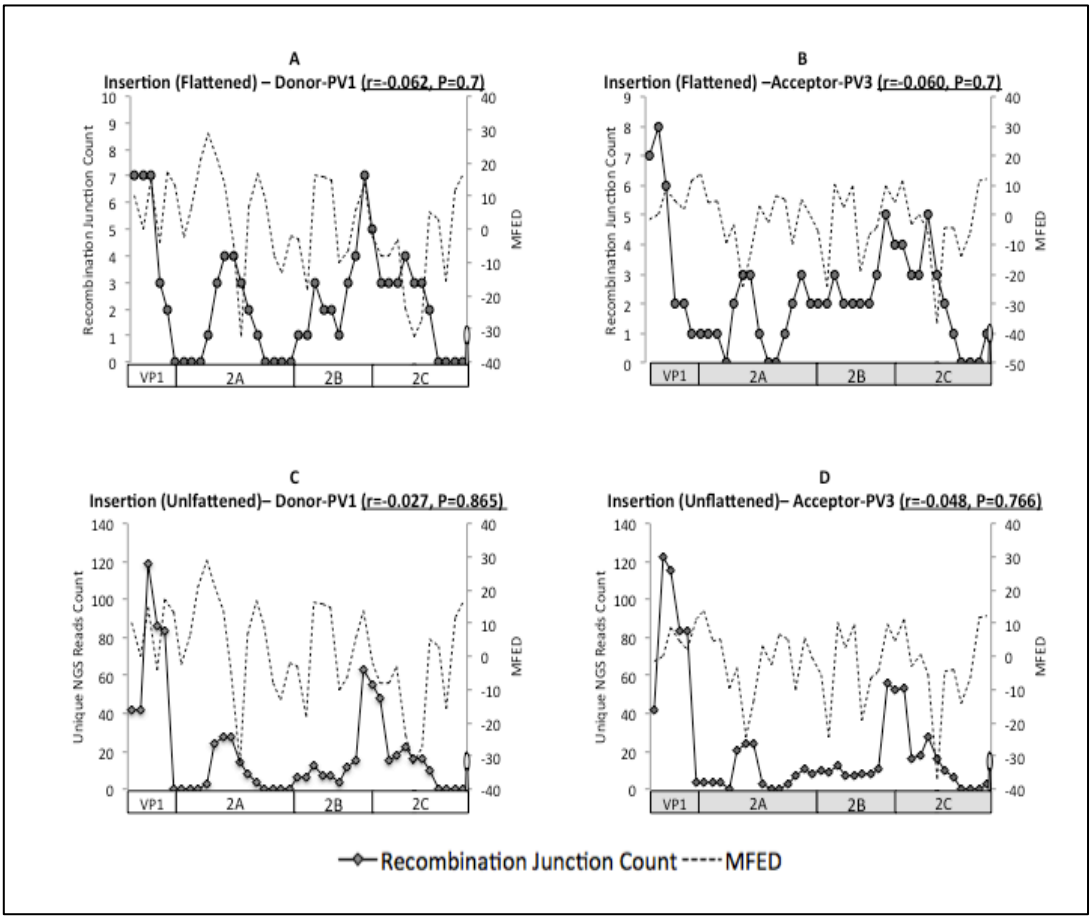


Figure 5-13 Imprecise-insertion Recombination RNA secondary structure analysis (flattened and unflattened data)

The mean folding energy difference (MFED) was measured within 100 nts sliding window (in 30 nts increments) in the sense strand of parental genomes indicated in dotted line and represented at the secondary y-axis. The solid line - marked with circles for flattened data and with diamonds for unflattened data - represents recombination junctions count within the same window represented by the primary y-axis (flattened) or the number of NGS reads (unflattened) . The x-axis was replaced by a targeted genome calibrated to the scale. White genome = PV1 (donor), grey genome = PV3 (acceptor). The title above each figure refers to the population name in its first half, and to the Sperman's Rank Correlation Test results in its second half.

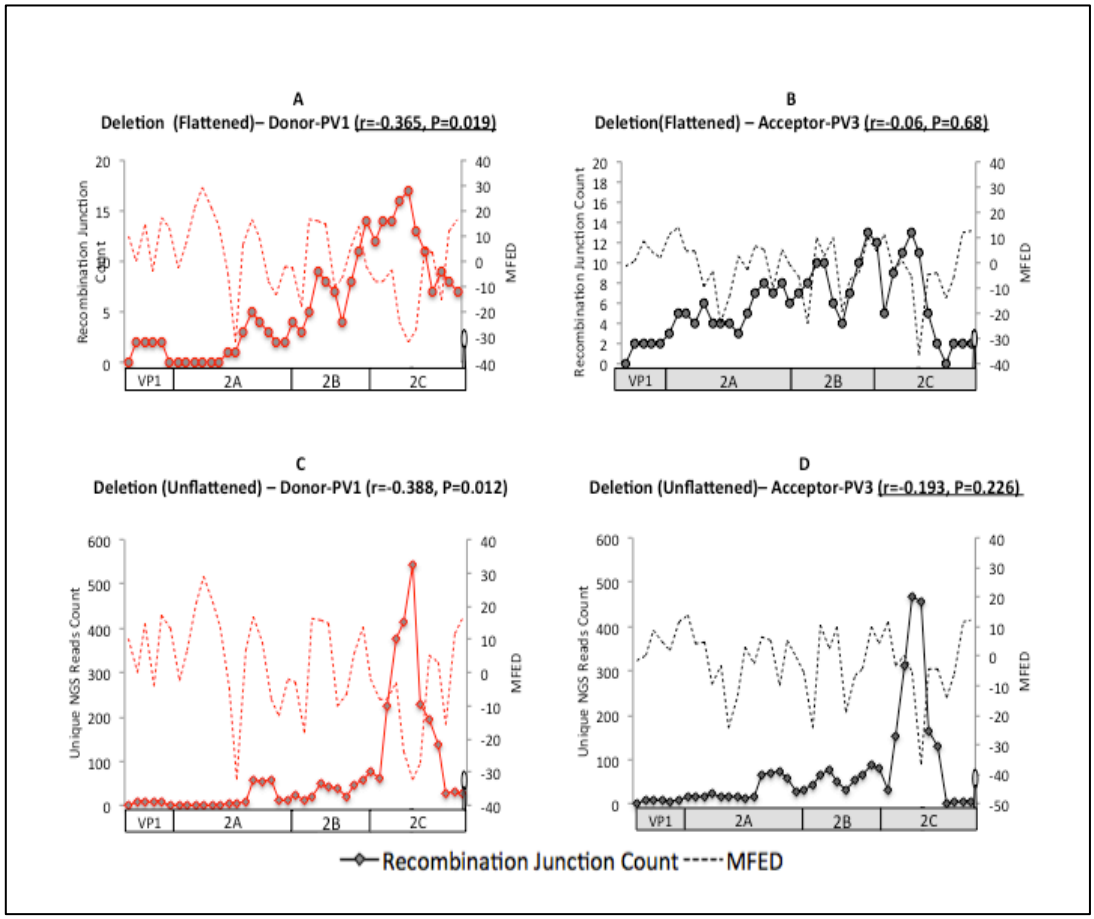


Figure 5-14 Imprecise-Deletion Recombination RNA secondary structure analysis (flattened and unflattened data)
The figure is highlighted red for the significant correlation findings. The remaining schematic features as in the previous figure.

5.9 Sequence similarities analysis of recombinants found by NGS

Sequence similarity between the parental genomes was suggested to be a key factor for recombination. This stems from the fact that the yield of intra-serotype recombinants was found to be 10 to 100-fold higher than intertypic recombinants (Molla et al., 1991, Lowry et al., 2014). To investigate this, the sequences similarities between the targeted regions of both viruses were measured within a 25 nts sliding window advanced by 1 nt (by Prof. David Evans), and then plotted against the number of observed recombination junction within the same window in the flattened data, and against the number of unique NGS reads in the unflattened data. The recombinants were divided into two samples, precise and imprecise. The imprecise recombinants were studied in both cases, flattened and unflattened data, while only flattened was studied for the precise recombination. The evidence of any correlation was determined by Spearman's test.

A very weak negative correlation was observed for the precise recombination ($r = -0.1$, $P<0.01$) in both genomes (Figure 5-15) PV1 and PV3 (represented as one line in the chart as the location is precise and would not change between the viruses). The correlation is very close to zero, which makes it negligible, and possibly it is an effect of the analysis method. This indicates a random distribution of the recombination sites over the genome with no significant preference towards regions with high similarities. Visually, there is no clear pattern among the precise recombinants (Figure 5-15). The line of recombination occurrences is moving up and down regardless of the sequence similarities. The same thing is awhen looking at the imprecise.

Previous studies suggested that imprecise recombination may exist adjacent to RNA similar segment between the two parental genomes (Kuge et al., 1986), or may not (Pilipenko et al., 1995). The results show no correlation with the similarity regarding the donor strand, and slightly stronger negative correlation than precise in terms of the acceptor strand in both cases, flattened ($r = -0.35$) and unflattened ($r = -0.25$) (Figure 5-16). This may indicate that imprecise recombination may favour the occurrence within divergent regions on the acceptor strand. Alternatively, as the correlation is not strong, similarities may not be a determinant factor for the

imprecise recombination, and the slight negative correlation with the acceptor strand may reflect the imprecise and random mechanism of the polymerase to re-associate, which is likely to end up on a divergent site.

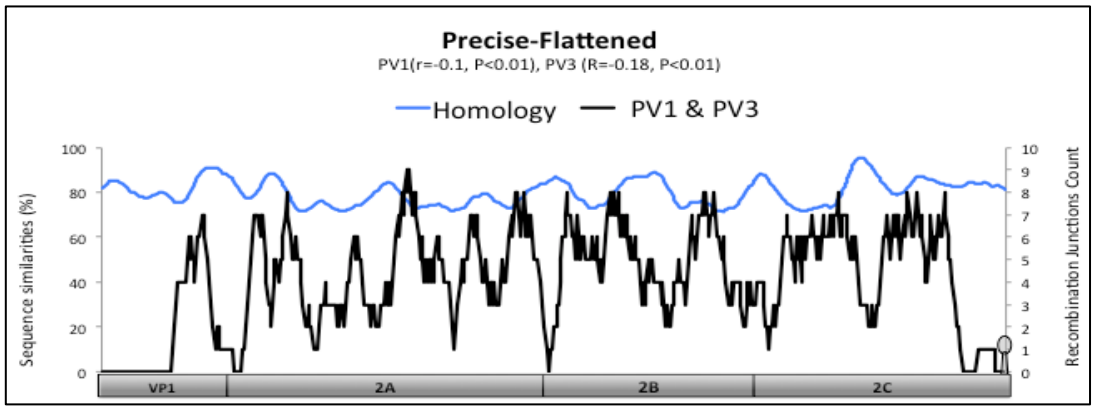


Figure 5-15 Sequence similarity analysis for precise recombination (flattened Data)
The sequence similarities (%) between the two viruses was measured within a 25 nts window advanced by 1 nts. (primary y-axis) represented by a solid blue line. The number of recombination events were counted within the same window (secondary y-axis). The the black line describes the relation between PV3 and PV1 and sequence similrities. The title contains the name of the population as well as the Spearman's Rank Correlation Test results for both viruses. The x-axis was replaced by the targeted genome with protein coding regions indicated. To reflect the correct location, it was calibrated to the scale. As each figure contains data for both PV1 and PV3, the colour of the genome is a gradient between grey (PV1) and black (PV3).

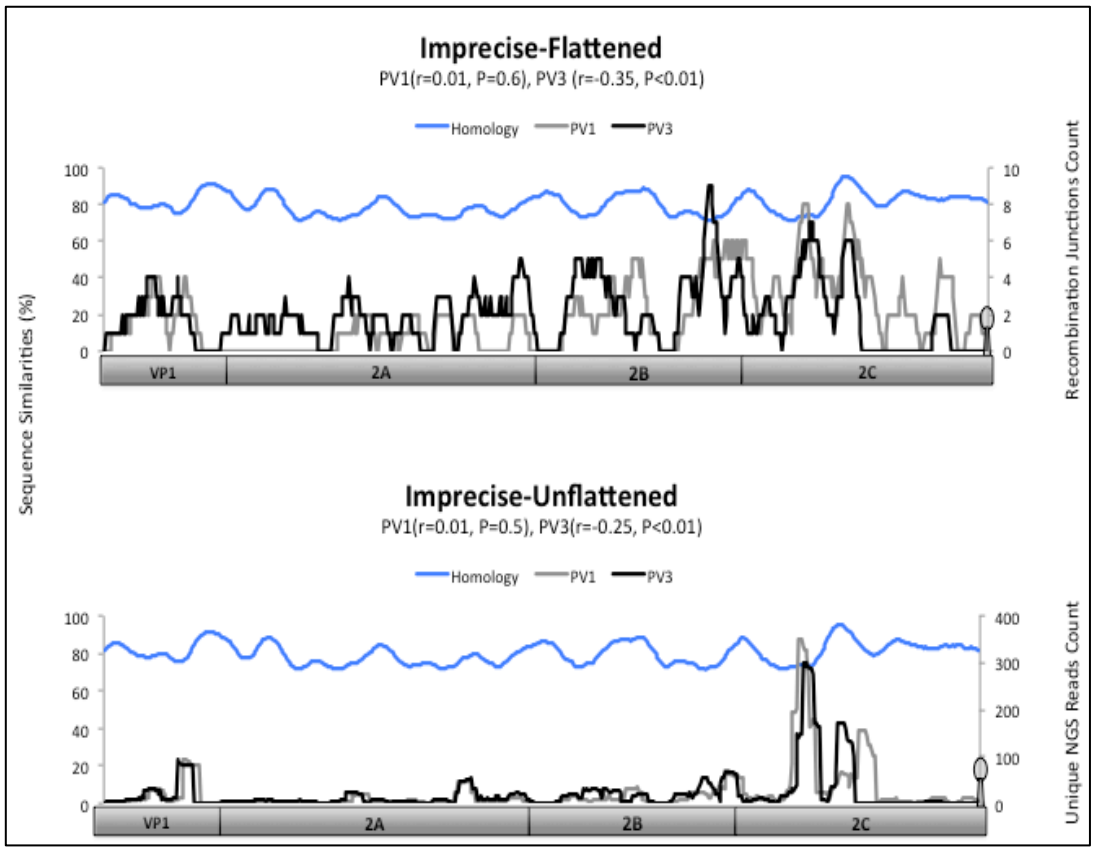


Figure 5-16 Sequence similarity analysis for imprecise recombination (flattened & unflattend data)
See Figure 5-15 for legend

5.10 Identical sequence analysis of recombinants found by NGS

The overall similarity between the two viruses (PV1 and PV3) does not seem to be a major influencing factor for recombination. To further investigate this, the correlation between the length of the identical regions and the occurrence of recombination was studied in flattened and unflattened data for both populations, precise and imprecise. Sequence identity between poliovirus type 1 and 3 is ~78% within the targeted region (P2), distributed as short identical sequences between 1 and 21 nts interspersed with variant nucleotides. The number of each identical length was counted across the targeted genome (Figure 5-17, Exp) and plotted against the observed recombinants at that particular length (Figure 5-17, Obs). Sequence identity of 1 nt length and the divergent sites were also included.

The results from precise recombination in Figure 5-17 show no significant difference between the length of identical sequence frequencies (1 – 21 mers) and the recombination junctions observed at these sequences (Figure 5-17). For example in the aligned sequences there were 98 different conserved dinucleotides, and 35 positions with an identity of 5 nts. Of the precise junctions characterised, 87 occurred at positions within identical dinucleotide and 32 occurred within identical sequences of 5 nts length. These findings go in line with results from previous studies (Lowry et al., 2014, Kirkegaard and Baltimore, 1986). However, looking at position 0 (divergent locations) between the two viruses, the observed precise recombinants are notably lower than the expected number.

On the other hand, the imprecise recombination does not show the same correlation. For example only 27 recombinants were observed within the identical sequence of 2 nts length, which is remarkably lower than the expected 89 recombinants. Nonetheless, it is also worth noting that imprecise recombination does not appear to favour the divergent sites either, as only a small number of recombinants were observed (24) in comparison to what was expected (228).

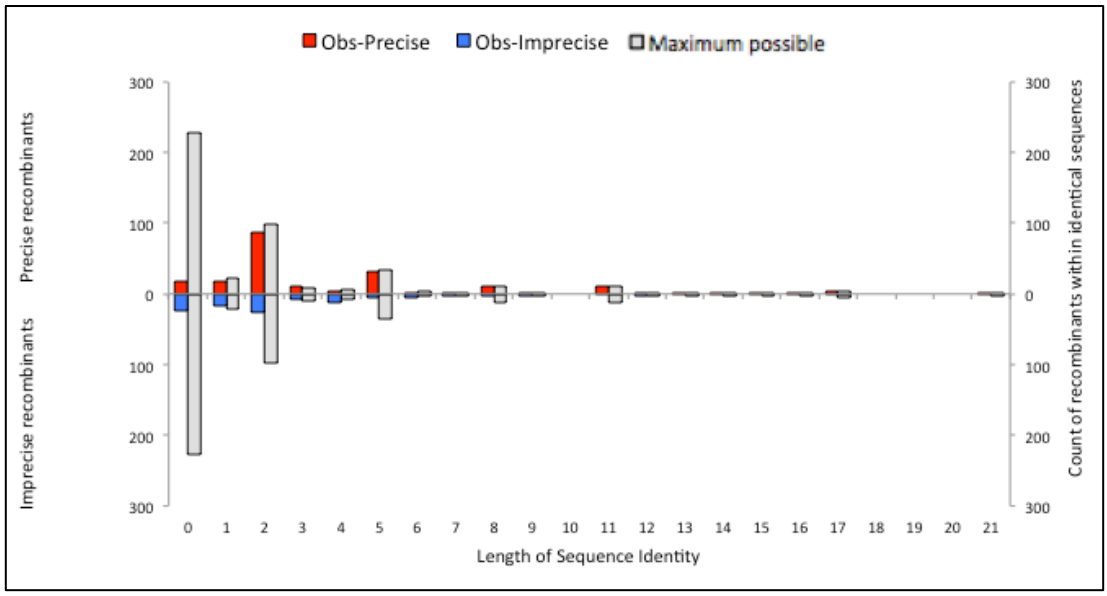


Figure 5-17 The occurrence of precise and imprecise recombination in regions of maximum sequence identity (flattened data)
Short identical sequences between 1 and 21 nts interspersed with variant nucleotides (x-axis), the 0 represents variant nucleotides. Individual counts of short lengths are represented by the grey bars (the primary y-axis). The red and blue bars represent the counts of recombinants found within the identical sequences of all possible lengths available in this region of the virus (the secondary y-axis).

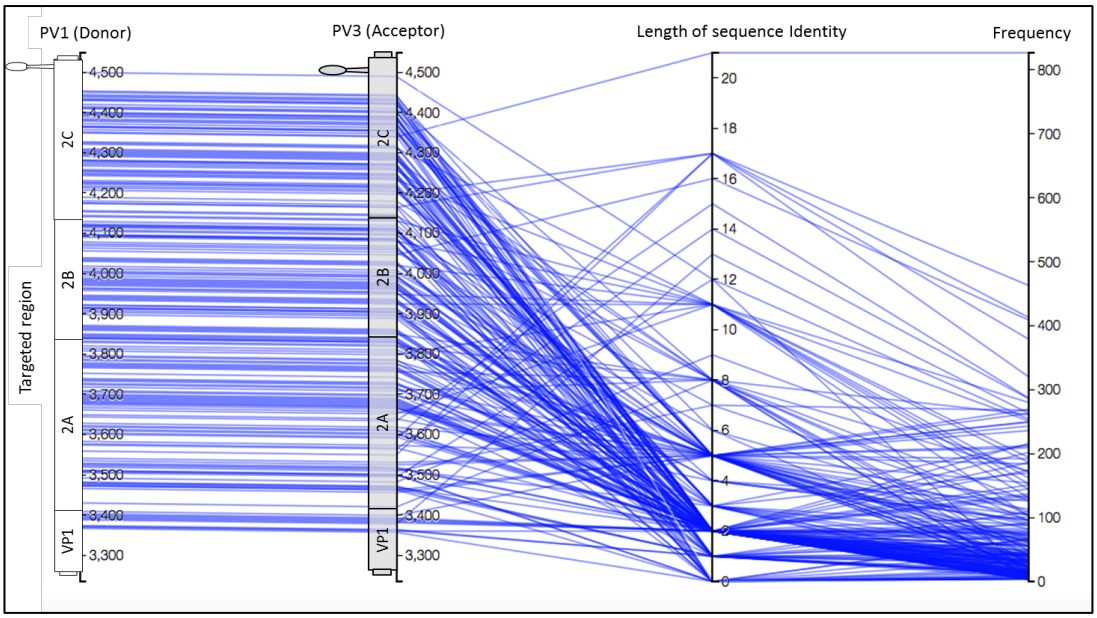


Figure 5-18 Correlation between the precise recombination frequences and the lengths of the identical sequences (unflattened data)
The same visualisation logic of the recombination map (**Figure 5-7**) is applied here. A vertical bar corresponding to the length of identical sequences was added. This links together the locations of each recombinant on the genome, the identical sequence within which it occurred, and the frequency. Blue lines = in-frame.

On the other hand, by analysing the correlation between the precise recombination from the unflattened data and the identical sequences length, it was found that the number of NGS reads corresponded to the length of the identical sequence rather than to a biological preference towards the identity (Figure 5-18). The longer the length of the identical sequences the higher the frequency. It is clear that the highest number of NGS reads was located at the identical sequence of 21 nts in the *2C*-encoding region, which occurs only once within the targeted region. Therefore recombinants at this region are represented by one line. This systematic bias may be caused by the inability to report the exact location of the junction (see 5.7.2 and 3.7.3). It is worth mentioning that the presentation of the recombination occurrences on the third axis (the length of sequence identity) in Figure 5-18 is actually a close-up look into the findings in Figure 5-17; the source of each recombinant occurring at a particular sequence identity length can be traced to the genomes.

In contrast, there seemed to be an inverse relationship in the imprecise recombination between the frequency and the sequence identity length. It could be observed that the number of NGS reads was reduced when the length of sequence identity increased (Figure 5-19). Interestingly, comparing the two maps together (Figure 5-18 and Figure 5-19), the region extends from the end of *VP1* to the first half of *2A* was found to be crowded with long identical sequence stretches and lack of imprecise recombinants occurrence. The crowdedness of identical sequences within this region (*VP1-2A*) is noticeable in Figure 5-18 as the blue lines (precise recombinants) in this region are pointing upwards towards several long identical sequences, 8,9,12,13,14,15. This observation possibly suggests that sequence identity has an important role in influencing recombination.

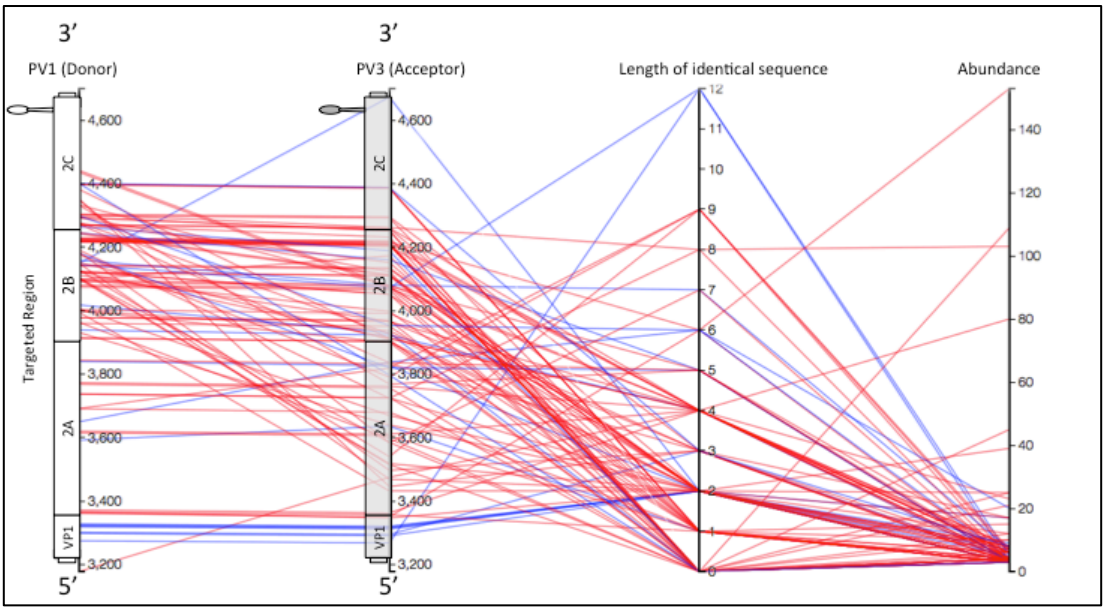


Figure 5-19 The frequency of imprecise recombination in regions of maximum sequence identity (unflattened data)
For the legand see Figure 5-18. This demonstrates imprecise-insertion and deletion together. Red lines = out-of-frame

5.11 Sequence Composition

The randomness, RNA structure and sequence similarities analysis did not show a robust involvement in determining the junction site. It was speculated that recombination occurs via a copy-choice mechanism in which RdRp paused for some reason during RNA synthesis and removes itself with the incomplete nascent RNA from the donor and resumes elongation on the acceptor (Kirkegaard and Baltimore, 1986, Nagy and Simon, 1997). Therefore, the primary objective of this section is to determine whether there were features of the nucleotide composition at the junction that might influence recombination.

Analysis of the sequence composition surrounding the junction sites in both parents was performed. These surrounding sequences were divided into four zones and are symbolised in a diagram in Figure 5-20 for precise recombination and Figure 5-21 for imprecise recombination. A specific name was assigned to each region based on its parental genome and its location regarding the junction site (Figure 5-20, A). For example R3 means the sequence located to the 3' side of the junction location on the recipient strand (PV3) and D5 corresponds to the sequence 5' to the junction on the donor strand (PV1). This naming system can be used to refer to specific nucleotide(s) within a particular region. For instance the last nucleotide transcribed by the polymerase is referred to as N-D3; the first nucleotide of the sequence located 3' to the junction site on the donor strand. Whilst the nucleotide where the polymerase re-associates is R5-N; the last nucleotide of the sequence located 5' to the junction site on the recipient strand. Based on this, the nucleotide upstream to R5-N is R5-N-1, whereas the nucleotide downstream to N-D3 would be N-D3+1. Using the same logic, NN-D3 would refer to the first dinucleotide of the sequence located 3' to the junction site on the donor strand.

The analysis method used is based on comparing the expected versus the observed for several sequence features. The recombinants were divided into two samples, precise and imprecise. Based on the fact that the number of NGS reads 'frequency' was found to be biased towards the long identical sequence stretches (see section 5.7.2) only the flattened data of the precise recombination was included in this analysis. In the case of imprecise recombination, the unflattened and flattened data

were considered for analysis. However, as the results did not show any difference, the flattened data only is explained in this section. Several Perl scripts (Appendix 5-9) were written to extract the desired mononucleotide, dinucleotide, homopolymer around the recombination junction by using the locations of every recombinant on both strands as reference points.

Five major analyses were performed:

1. G+C and A+T content within 50 nts of each sequence around the junction. These sequences are referred to as D3, D5, R3, R5 in Figure 5-20, A.
2. Mononucleotides. This was divided into two analyses; mononucleotides within a 50 nts window of the surrounding sequences and mononucleotides at the junction site. The nucleotides analysed by the latter are illustrated in Figure 5-20, B. They are highlighted either red or blue depending on which genome they belong to. They have different names indicating that they were analysed individually.
3. Dinucleotides at the junction site. These targeted dinucleotides are demonstrated in Figure 5-20, C.
4. Homopolymer within a window of 50 nts around the junction location.
5. Exit and entry nucleotides. The exit nucleotide (referred to as Z) is the one upstream to where the polymerase dissociates from the donor strand while the entry nucleotide (referred to as Y) is where the polymerase re-associates on the acceptor strand. They are highlighted in blue and red respectively in Figure 5-20, D.

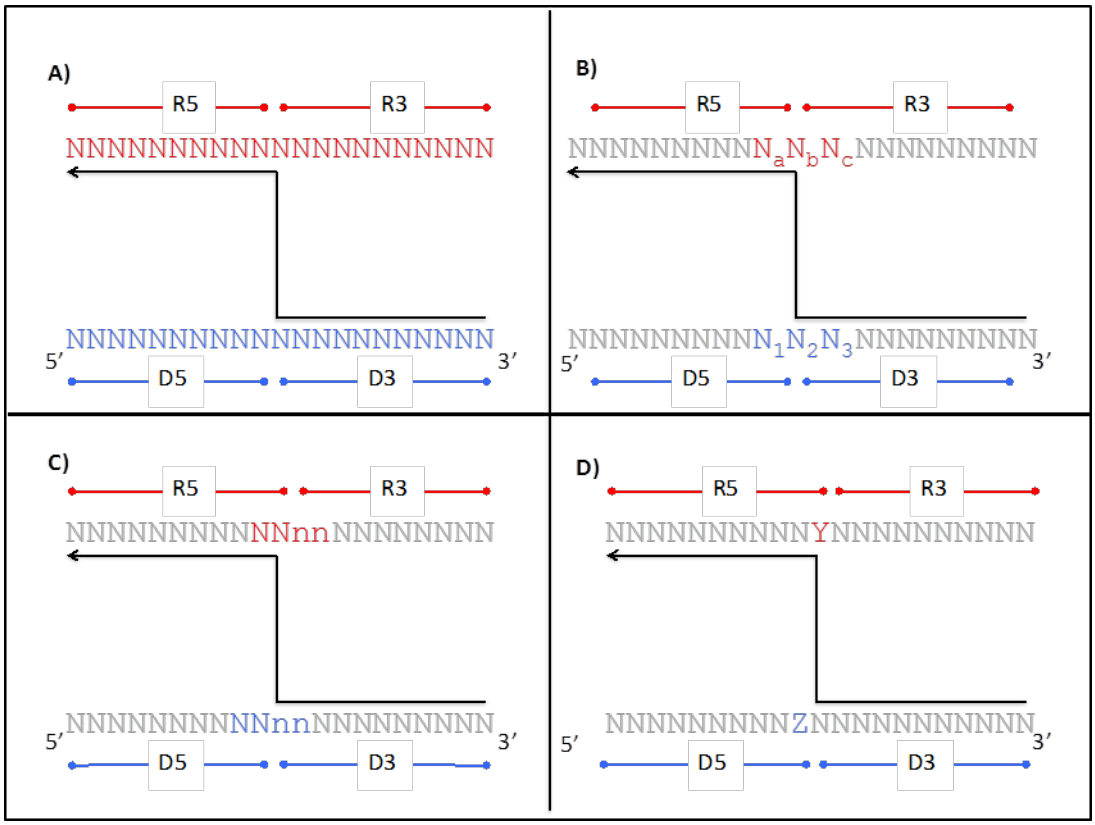


Figure 5-20 Sequences surround the junction site considered for sequence composition analysis (precise recombination)

The sequences that encompass the junction location were divided into four zones and named based on their location to the junction. R5 = 5' end to the junction on recipient strand, R3= 3' end to the junction on recipient strand, D5 = 5' end to the junction on donor strand, and D3 = 3' end to the junction on donor strand. The N's refer to any possible nucleotide. Red refers to the donor strand and blue to the recipient strand. The black arrow denotes the path of the RdRp A) General overview of the four zones. B) Demonstration of the nucleotides that were nominated for the mononucleotide test. C) The dinucleotide nominated for the dinucleotide test. D) Defining the entry and the exit sites for the RdRp, Z = Exist site, Y = Entry site. The grey N's refer to nucleotides located at a far distance from the junction, and were thus excluded from this analysis.

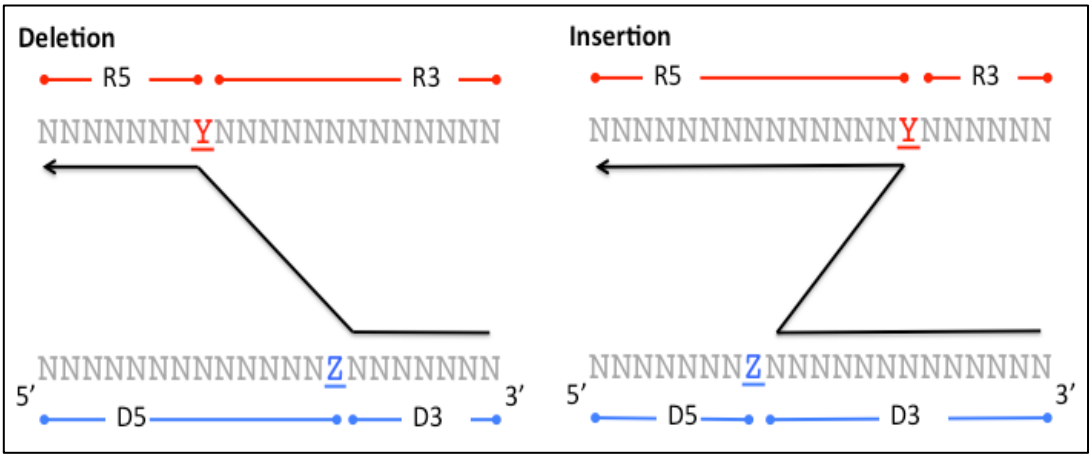


Figure 5-21 Sequences surround the junction site considered for sequence composition analysis (Imprecise recombination)

The same type of nomenclature used in the precise recombination is applicable on the imprecise recombination for all the analyses. This figure illustrates the entry and the exit sites analysis in the imprecise recombination as an example.

5.11.1 G+C and A+T content (GC-Content, AT-content)

It was reported that G+C and A+C content are associated with bias in the recombination mechanism (Runckel et al., 2013). The rational is that if the RNA incomplete strand dissociated at G+C-rich sequences, it could be predicted to anneal to the acceptor strand more robustly than A+U-rich sequences. To investigate this, a comparison between the frequency of G+C and A+T within 50 nts window from the four zones (D3, D5, R3, R5) with the frequency of G+C and A+T in the targeted region (the region between nucleotides 3235 and 4548 in the virus genome) was performed. The expected ratio was calculated as (G+C/A+T+G+C)*100 for GC and (A+T/A+T+G+C)*100 for AT. The denominator equals 1300, which is the length of the targeted sequence. The GC-content – within this region 3235-4548 - was found to be 45.3% and 44.6% for donor and acceptor respectively and 54.6% and 55.3% for AT-contents. This was compared with the observed, which was calculated by calculating the GC-content within a 50 nts window as (G+C/50)*100 for each region. The statistical significance was tested by Chi-Square. None of the 50 nt sequences flanking the recombination junction in the donor or acceptor had shown any bias with either content in both populations, precise and imprecise (Figure 5-22).

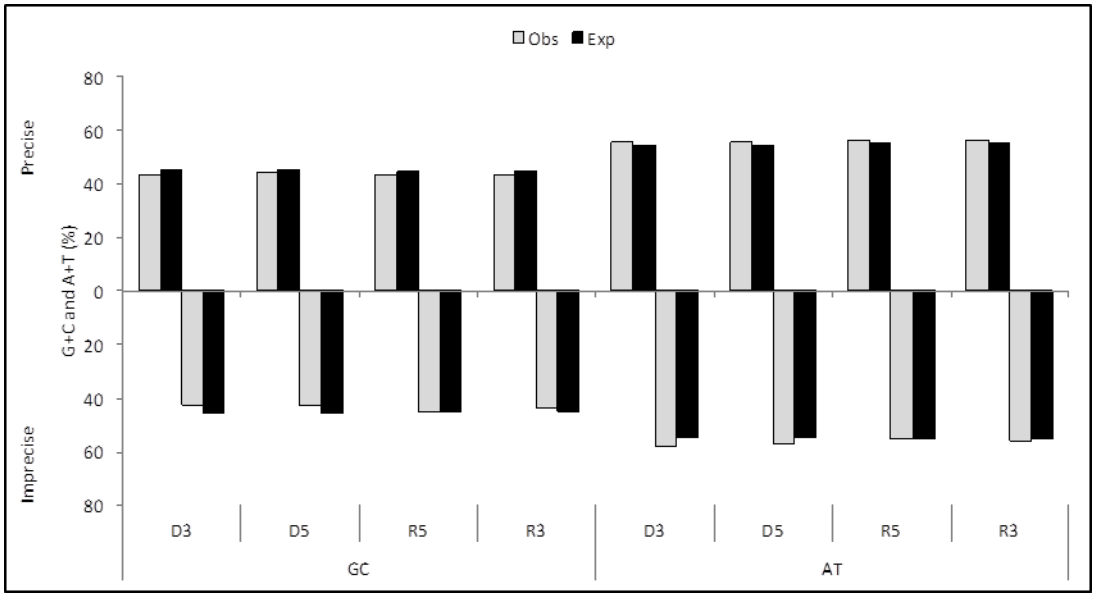


Figure 5-22 G+C and A+T contents analysis (within 50 nts window) in precise and imprecise recombination
The frequencies of the expected GC or AT contents (black bars) were binned side by side with the observed frequencies (grey bars) in all the four zones for both populations, precise and imprecise. Y-axis represents the frequency (%) for GC and AT contents; the upper side belongs to precise recombination, and the lower one for imprecise recombination. X-axis shows the test in question (GC or AT) and the zone within which they were analysed.

5.11.2 Mononucleotides

The high number of distinct recombinants identified in this project implies that, if a specific nucleotide sequence is required for recombination, it must be a very short one. It was speculated that A-rich sequences would influence RdRp slippage facilitating its dissociation (Pilipenko et al., 1995, King, 1988). To inspect if there is any nucleotide(s) bias, mono and dinucleotide were analysed. In regards to mononucleotides the test was split into two analyses. First, a comparison between each nucleotide frequency within 50 nts of the four zones (D3, D5, R3, R5) with the frequency of each nucleotide within the targeted region of the genome (P2 region). The same method as in GC-content test was used to calculate the frequency; no evidence for any bias was detected (Figure 5-23).

Secondly, a study of the mononucleotides at the junction site was carried out. To do this, three nominated nucleotides were picked from each strand; these represent the nucleotide where the junction occurs (N-D3 and R5-N), the nucleotide upstream (N-D3-1 and R5-N-1) and the nucleotide downstream (N-D3+1 and R5-N+1). The rational of picking these particular three nucleotides is to study if there is any preferred nucleotide for recombination either at the junction site or at the border. For the sake of simplicity in demonstrating the findings in a graph they were given different names. Those corresponding to donor strand are N_1, N_2 and N_3, while for the acceptor strand they were called N_a, N_b, N_c Figure 5-20, B. The comparison is based on the difference between the number of observed recombinants with a certain nucleotide and the frequency of that nucleotide in either parent (expected). For example, nucleotide ‘G’ was found with 23% frequency in the donor strand targeted region, thus if the recombination is random we expect to see ~23 recombinants of 100 have ‘G’ at their N1, N2 or N3 positions. The statistical significance was tested by the Binomial Distribution test.

The results (Figure 5-24) show some noticeable biases; nucleotide ‘C’ was found to be selected against in both populations (precise and imprecise) at N2 site where the RdRp dissociates. Additionally, ‘A’ nucleotide was favoured at N_2, N_3 and N_C. It is also worth noting that ‘T’ nucleotide was found lower than expected at two sites in the imprecise recombination (N_2, N_3), which probably may suggest a significant role for this nucleotide in the imprecise recombination.

The different biases observed between the two populations may reflect the presence of two different mechanisms for imprecise and precise. Nucleotides N_C and N_3 correspond to each other by their locations (right before the dissociation and re-association sites) were found to be significantly high among the imprecise recombinants. Interestingly, separating the imprecise into insertion and deletions (Figure 5-25) revealed the presence of such correlation in the imprecise-insertion only. This suggests that the mechanisms underlining imprecise-insertion and –deletion may differ from each other. Nevertheless, as there is more than one pattern, it is difficult to conclude from these findings alone how these nucleotides correlate and interact with each other. These patterns can be verified by repeating the experiments and by considering some more experiments including sequence manipulation.

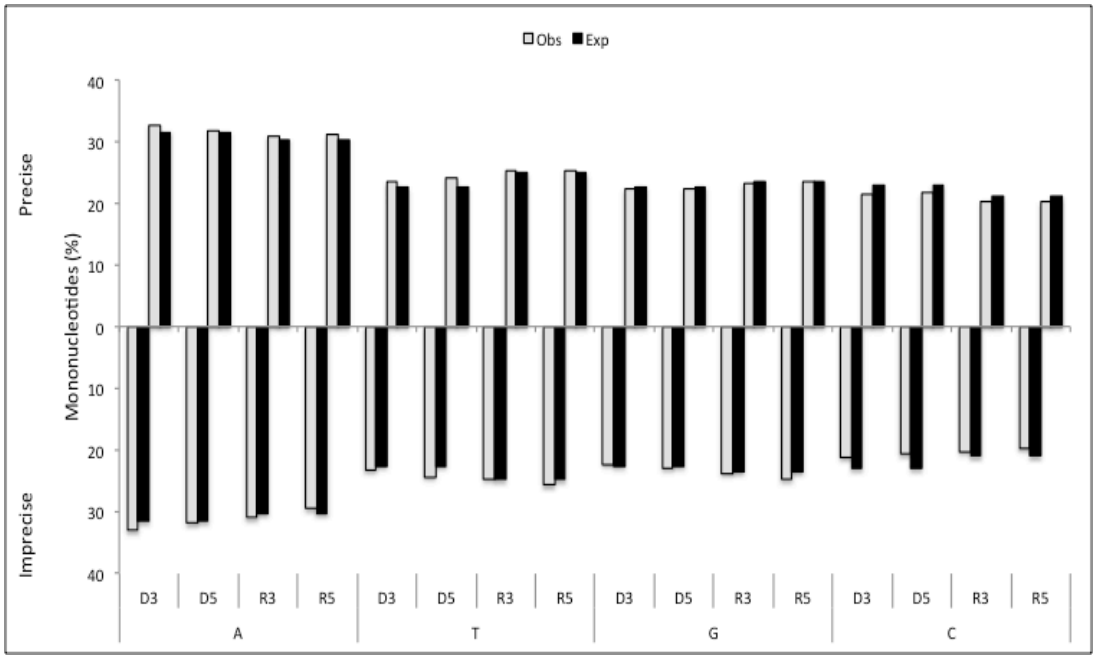


Figure 5-23 Mononucleotide analysis (within 50 nts window) in precise and imprecise recombination
For the legends see Figure 5-22

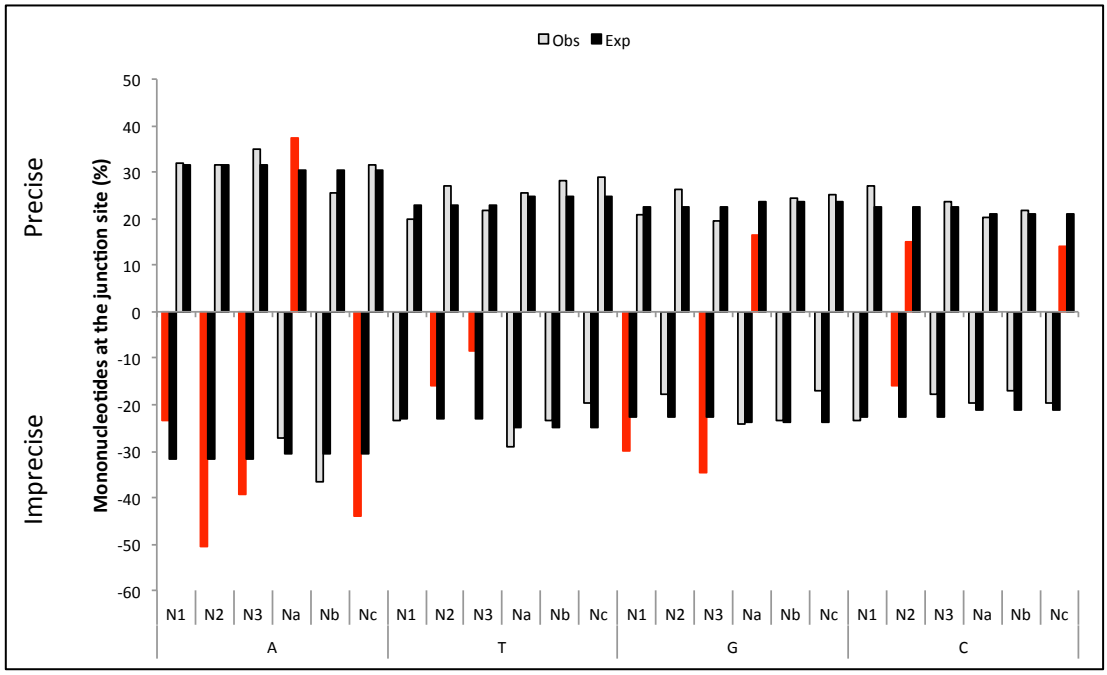


Figure 5-24 Mononucleotide analysis (at the junction site) in precise and imprecise recombination.
The six nucleotides which were picked for this analysis are illustrated in **Figure 5-20**, B. N1, N2, N3 are located on the donor strand while Na, Nb, Nc are located in the recipient strand. The significant difference between the observed and expected is highlighted in red (calculated by Binomial Distributing test). X-axis demonstrates the mononucleotide in question and its corresponding name as defined in in **Figure 5-20**, B.

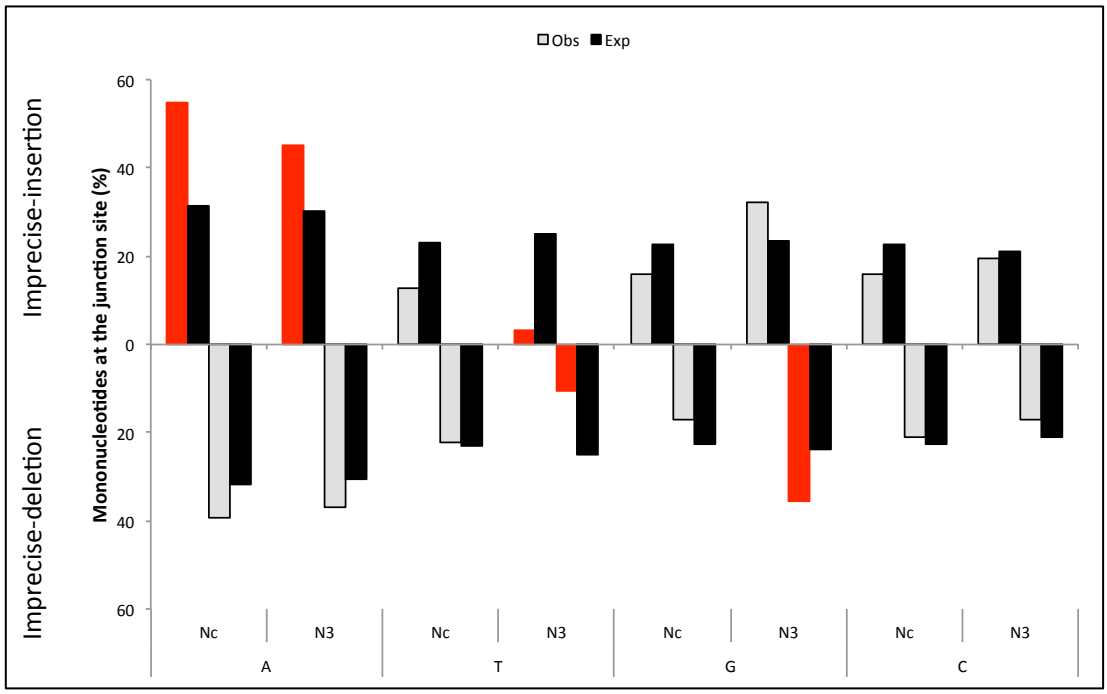


Figure 5-25 Comparative analysis of mononucleotide at the junction site between imprecise insertion and deletion.
Analysis of mononucleotides Nc and N3 from **Figure 5-20**, B at the junction site. The Nc is located on the acceptor and the N3 on the donor strand. The significant difference between the observed and expected is highlighted in red (calculated by Binomial Distributing test). X-axis demonstrates the mononucleotides and their corresponding names as defined in in **Figure 5-20**, B.

5.11.3 Dinucleotides

Dinucleotide 'AA' was noticed by King to be significantly higher than the random model in the area of the junction site (King, 1988). To further investigate this and any other bias towards any of the potential dinucleotides, the same method of studying the mononucleotides at the junction site was used, with the focus on dinucleotide this time. The locations of the dinucleotides in regards to the junction site are illustrated in Figure 5-20, C. Unlike the mononucleotide around the junctions where six different nucleotides were tested and every nucleotide was given a specific name, dinucleotide exists with one possibility at the junction site, and therefore they are referred to by the zone's name.

The frequency of all the 16 dinucleotides was determined in the targeted genome individually. The expected values were calculated by the same method used in the mononulcutide (see previous section). The four zones around the junction were tested and the statistical significance was measured by the Binomial Distribution test.

The results (Figure 5-26) of all the four zones showed some desirable dinucleotide for the recombination mechanism. As for the precise recombination, several dinucleotide patterns can be observed at the four zones. For example, 'GA' occurred significantly higher than expected at the R3 and D3 zones, which are corresponding to each other in terms of location. Similarly, 'TC' was found to be significantly high at the R5 and D5. It could be speculated that this arrangement might play a role in facilitating the recombination by aligning the two parental viruses. However, it is difficult to interpret the bias found as precise recombination occurred majorly within identical sequences.

On the other hand, different other significant patterns were found among the imprecise recombination. In consistent with previous studies (King, 1988, Nagy and Simon, 1997), dinuclutides 'AA' was found to be significantly rich at the dissociation site (D3 zone). Moreover, another dipurines – such as 'AG' and 'GG'- were found to be significantly high at this site, which is in agreement with previous results found in Evans lab. More biases were observed at different zones for the imprecise recombination, which might also play a role in facilitating recombination (see Figure 5-26).

5.11.4 Homopolymer (50 nts window)

In addition to looking for potential biases in mono or di-nucleotide distributions and recombination junctions the presence of short homopolymeric sequences was investigated. To study whether homopolymer is a recombination influencing factor, all possible homopymers of 3-6nts were identified in the targeted genome in both genomes (donor and acceptor) and compared to the observed homopolymers within 50nts sequences of all the zones (D3, D5, R3, R5). As no correlation was found in any of them, only the results of the 3nts hompolymers were demonstrated in this chapter (Figure 5-27).

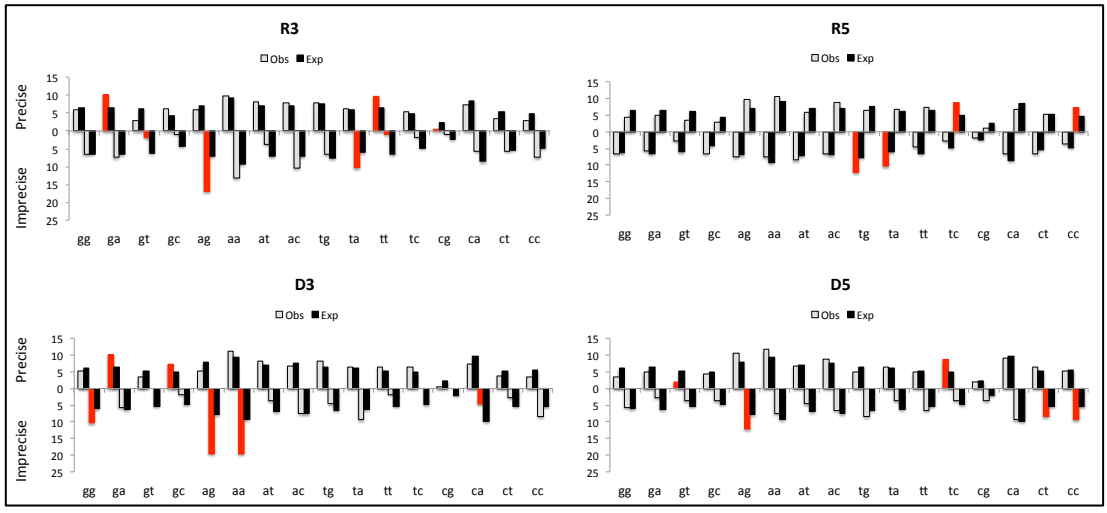


Figure 5-26 Dinucleotide analysis (at the junction site) in precise and imprecise recombination for zones R5 and D3
The upper panel belongs to R5 zones and the lower for D3. X-axis indicates the dinucleotide in question. See **Figure 5-24** for the rest of the legends

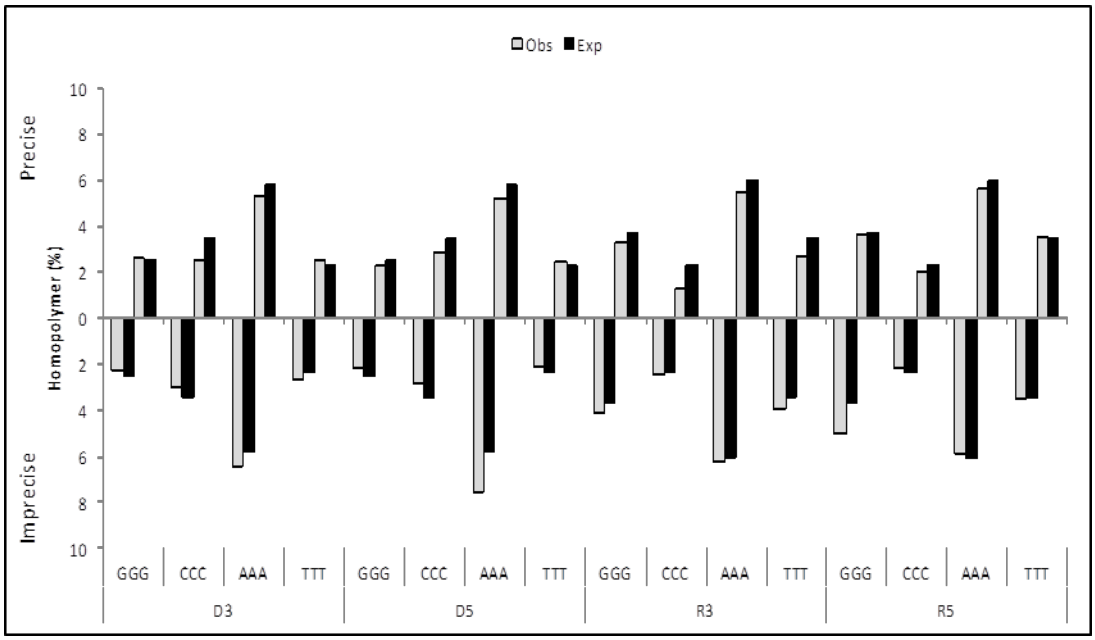


Figure 5-27 Homopolymer analysis (within 50 nts window) in precise and imprecise recombination
The figure demonstrates the analysis of 3mer homopolymers. X-axis demonstrates the homopolymer in question with its corresponding zone. The rest of the variables as in previous figures of section 5.11.

5.11.5 Sequence compositions of the identical regions

The analysis of precise recombination sequence compositions within 50 nts of the surrounding sequence at the junction location did not show any significant role of sequence in influencing recombination. To further study this, the sequence compositions within the shared identical sequences between the two parents at the recombination site were analysed in the case of precise recombinants (see Figure 5-28). The recombinants within the precise recombination population were divided into sub-samples based on the length of the identical sequences near the junction location. Subsequently, the sequence composition in question was measured for each recombinant in a particular sub-sample and the average was calculated. Finally the average of all the averages from the sub-samples was calculated and compared with the expected values - for both parents - using Chi-Square.

The expected values were calculated in terms of the targeted sequence length; 1300 nts - the region between nucleotides 3235 and 4548 in the virus genome. For example, the equation '(G+C/1300)*100' was used to calculate the expected ratio for GC. Additionally, the calculation of the percentage of the sequence composition was modified according to the length of the identical sequence in question. For example, the dinucleotide 'AT' would constitute 100% of the sequence if the identical sequence was composed of two nucleotides and happened to be 'AT'. However, this will decrease to 50% if the AT was found within an identical sequence of three nucleotides and 33.3% within four nucleotides and so on. Custom Perl codes were used to calculate the percentages of sequence compositions (appendix 10-13).

The results of the analysis (Figure 5-29) did not show any significant correlation in all the studied sequence compositions. This would increase the importance of the sequence identity rather than the sequence composition in influencing precise recombination.

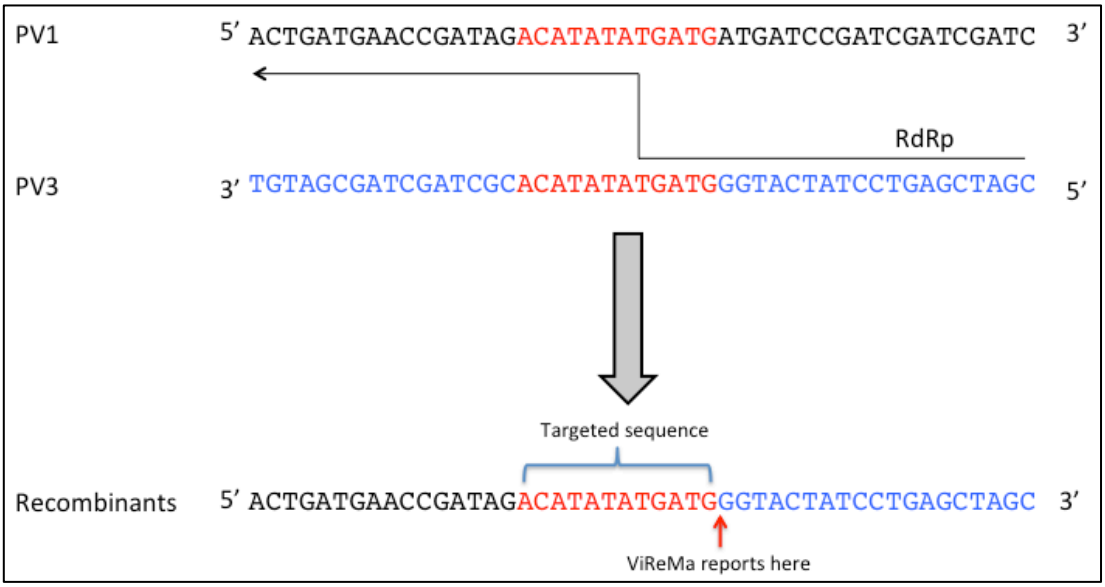
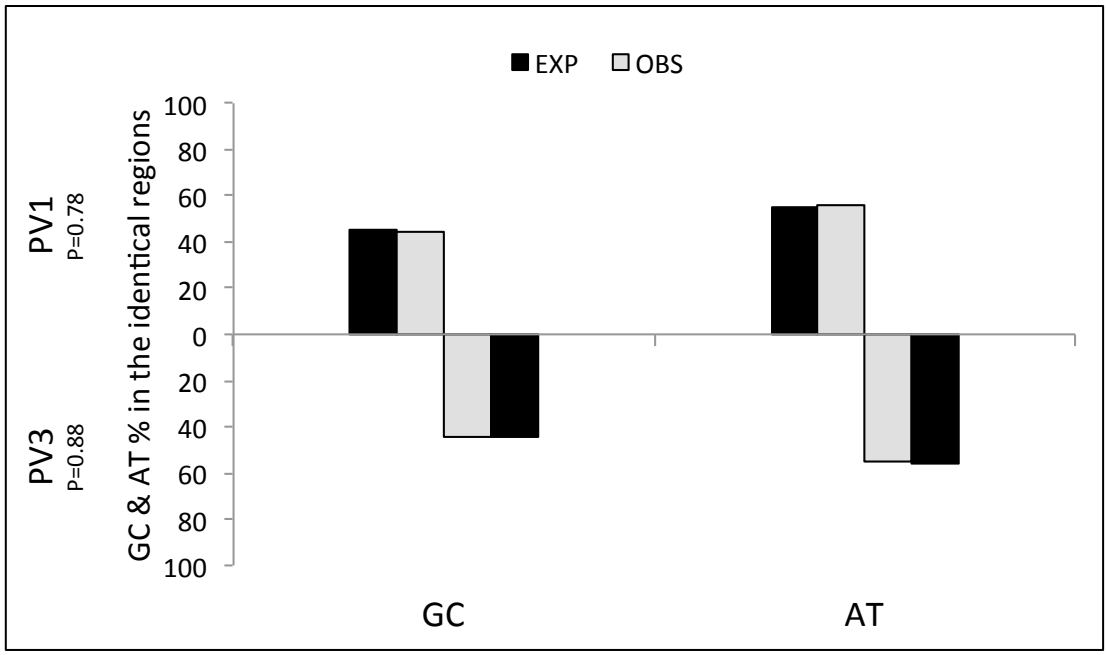


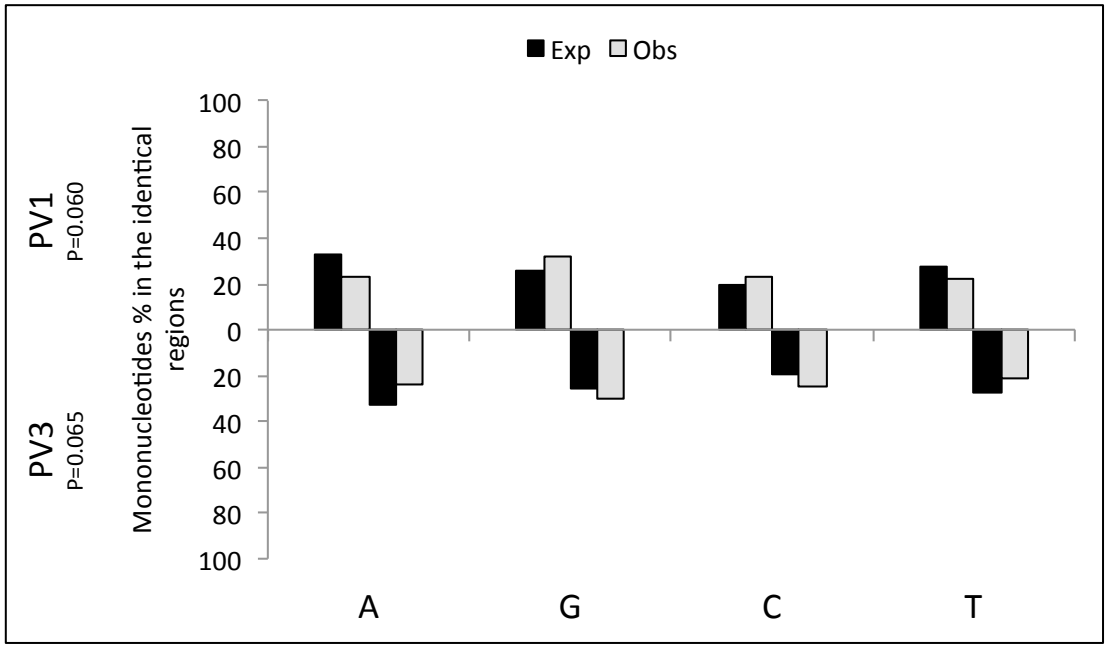
Figure 5-28 Demonstration of the identical sequences picked for sequence composition analysis in precise recombination

The figure demonstrates an example of an identical sequence of 12 nts, within which the recombination was occurred; the sequence represented in red. The black and blue represent PV1 and PV3 viruses respectively. ViReMa (depending on the configurations of the algorithm) would report the resulted recombinants at the 3' of the identical sequence (the red arrow), which means that the red sequence would be considered as a PV1 sequence. Based on this, the Perl codes were written to extract the red sequence and analyse its sequence composition.

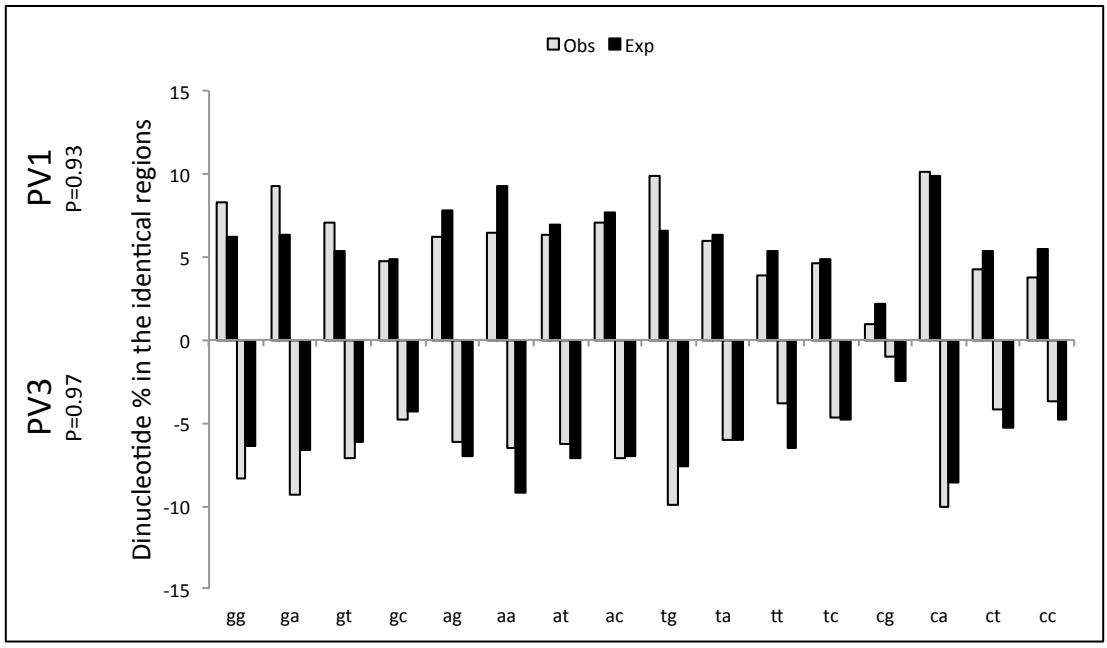
A)


EXP
OBS
PV1
P=0.78
PV3
P=0.88
GC & AT % in the identical regions
100
80
60
40
20
0
20
40
60
80
100
GC
AT

B)


Exp
Obs
PV1
P=0.060
PV3
P=0.065
Mononucleotides % in the identical regions
100
80
60
40
20
0
20
40
60
80
100
A
G
C
T

C)



D)

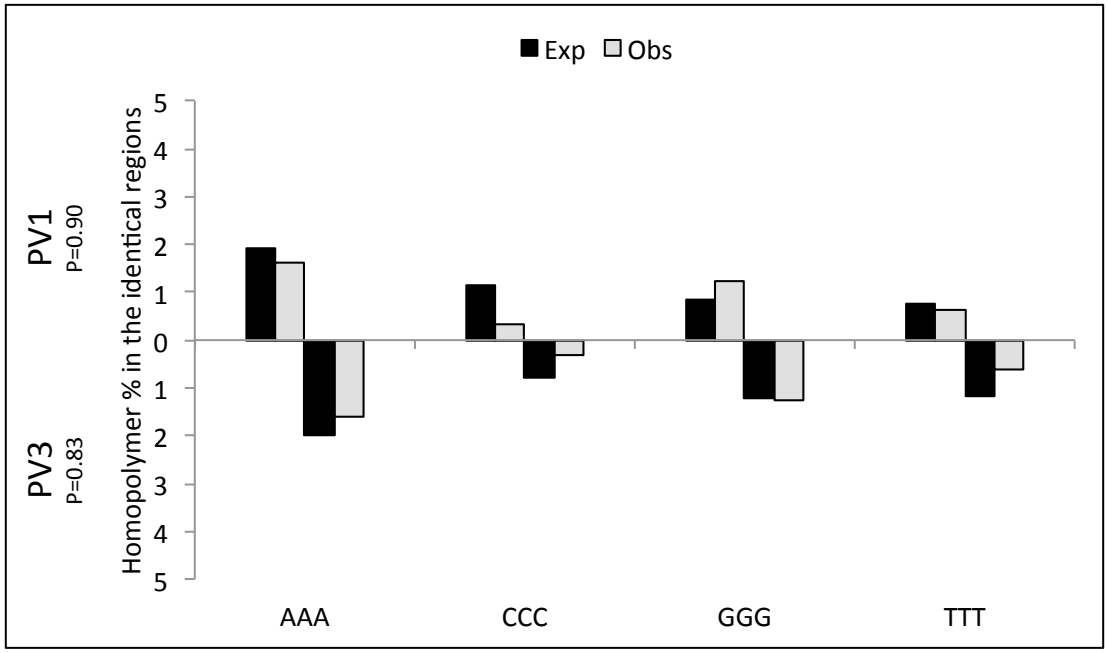


Figure 5-29 Analysis of sequence composition within the identical sequences
The expected and the observed values for each sequence composition were compared in both viruses; represented in the figures as PV1 at the upper side and the PV3 at the lower side of the charts. The chi-square p value is written for each test near the y-axis and below the viruses' names. The grey bars denote the observed values and the black for the expected values. A) Analysis of GC and AT contents. B) Analysis of the mononucleotides. C) Analysis of the dinucleotides. D) Analysis of the homopolymers.

5.11.6 Exit (Z) and entry (Y) nucleotides

In the majority of the identified recombinants, it was not possible to unambiguously define the recombination junction due to local sequence identity between the two viruses. In this report, the junction was defined at the 5' of the identical sequence (see section 5.3). In this test we examined the nucleotides combination between the Z and Y nucleotides (Figure 5-20, D and Figure 5-21) and compared them to the expected nucleotides combination. The expected ratio was calculated by multiplying the frequency of each nucleotide at the Z location in the donor strand by the frequency of each nucleotide at the Y location in the acceptor strand. For example the frequency of nucleotide 'A' within the targeted region in the donor strand was found to be 32%, and 'G' in the acceptor strand was 24%. Thus, the expected value equalled 0.32 * 0.24 = 7%. If the combination of Z and Y is random, then we expect to see 7% of the recombinants with 'A' as an exit site and 'G' as an entry site.

The results showed a significant increase in the transitions and decrees in the transversions of the precise recombination. To further investigate this, 90 nts from both sides of the precise recombinants were extracted. The analysis was done by Professor David Evans; the reading frame (RF) was divided into three possibilities:

1. RF1 = `.ACG.TAG.CTA.acg.tag.cta.`
2. RF2 = `AC.GGA.TCG.Aac.gga.tgc.a`
3. RF3 = `A.CGG.ATT.ACa.tta.tat.gt`

All the recombinants were then translated using Manipulation Suite (bioinformatics.org). If the recombinant had no stop codons, it meant that they were in frame, since there were 90nts before and after the recombination junction. Therefore those recombinants were categorized under the RF1. For anything that had stop codons, one or two nucleotides were added to the 5' end of the recombinants and checked whether it restored the open reading frame. If 1 nt was needed then the recombinant would be categorized under RF2, while if 2 nts was needed then it was RF3.

The results in Figure 5-31, showed that RF1 existed with ~84%, while RF2 and RF3 were ~8 % and ~6% respectively. These findings suggest that the large number of transitions in precise junctions was in fact due to the way that the junction location was defined; at the 5' of the identical sequence. The vast majority of the recombinants were in RF1, reflecting that Z and Y nucleotides were occupying the third base of the codon immediately preceding the recombination junction (Figure 5-20, D), therefore, transitions were more common. This is because in previous studies, transitions were found to outnumber transversions (Kuge et al., 1989, Acevedo et al., 2014). Additionally, it was found that only 32% of the nucleotides at the third positions of the codons were conserved between poliovirus serotypes and around 98% at the first and second positions (Toyoda et al., 1984). This was reflected by the difference between RF1 and RF2 and RF3 found in this analysis. Due to the inability to precisely define the junction location, this was not pursued any further.

On the other hand, studying the exit and entry sites among the imprecise recombination, a bias towards 'A' nucleotide as an entry site was observed (Figure 5-30). This probably reflects the percentage of nucleotide 'A' within the targeted region, which is ~31% in comparison to around ~22% for 'C', 'G', 'T' individually. Nonetheless, this could be verified by investigating more data.

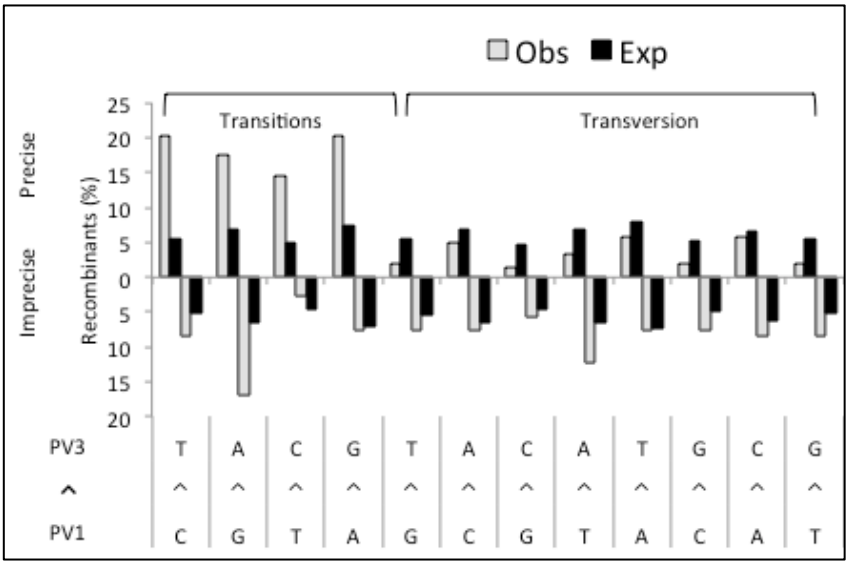


Figure 5-30 Exit and Entry nucleotides analysis
The combinations of Z and Y nucleotides in 'Figure 5 20, D' were tested by comparing the expected (black bins) with the observed (grey bins). The x-axis demonstrates the possible combinations that can happen when Z changes to Y, divided into transitions and transversions for the precise recombination. The nucleotides below the arrows represent nucleotides from the donor strand. Above the arrows are nucleotides from the acceptor strands. (^) = 'change to'. The y-axis represents the count of recombinants (%) found for each combination.

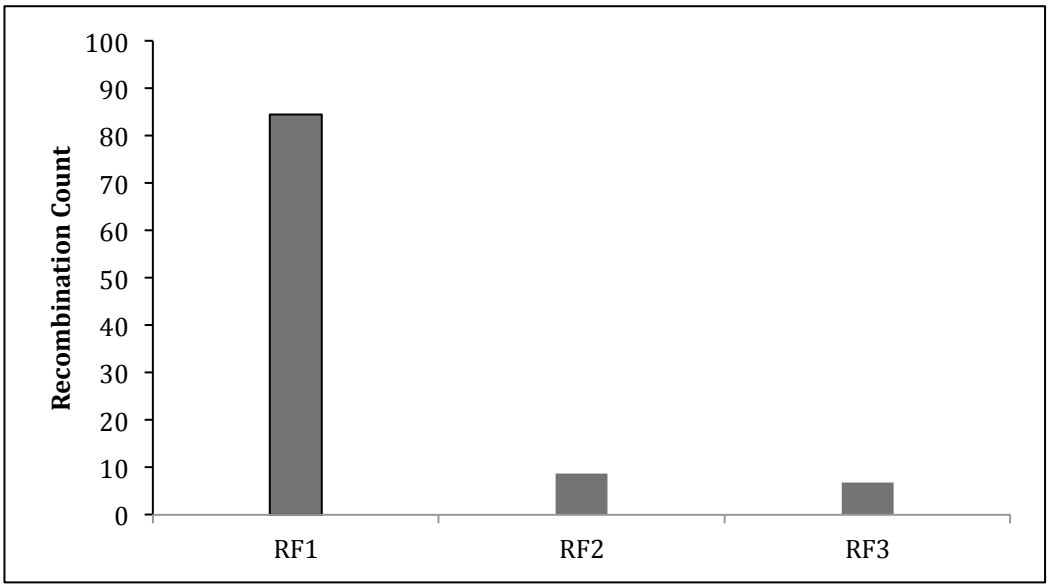


Figure 5-31 Reading Frame (RF) in Precise Recombination
The 206 precise recombinants were tested for their RFs to determine at which position of the codon the recombination occurred**.** The x-axis represents the possible RF, which described in 5.11.6. The y-axis refers to the percentage of recombinants found for each possible reading frame.

5.12 Discussion

In order to understand the mechanism that influences poliovirus recombination, an experimental system that can specifically amplify recombinant molecules from a wild type mixed infection was developed. This was accompanied by NGS technology to sequence and analyse recombinant RNA that was present 5h post infection before the virions were produced. Therefore, minimising the selection for increased fitness within the recombination population that on its turn would widen the range of obtained recombinants. To analyse the data, a bioinformatics system was created and optimised, through which high-resolution recombination maps were produced for both precise and imprecise recombinants (Figure 5-7).

Using the bioinformatics pipeline to analyse the NGS reads, the total number of recombination junctions was found to be 313 distinct recombinants. Of those 206 were precise and 107 were imprecise recombinants. Of the latter, 35 were imprecise-insertions and 70 were imprecise-deletions. The results demonstrated that at this early stage of the infection and within an environment with minimised selection for fitness, precise and imprecise recombinants are detectable in the infected cells. In contrast to the results obtained by Jarvis & Kirkegaard, although they harvested the virus progeny early in the infection and the experimental system they used did not select for viability, no imprecise recombinants were detected (Jarvis and Kirkegaard, 1992). This discrepancy in results can be attributed to two reasons. Firstly, they used a smaller infection scale, which might have undermined the presence of the imprecise recombination. Secondly, they studied intratypic recombination within which parental genomes are identical, which might have disfavoured the occurrence of imprecise recombination. In fact, in this study, imprecise recombination was found to be selected against in the presence of long identical sequences. This was reflected by the absence of imprecise recombinants within the region extended from the end of *VP1*-coding region to the middle of 2A-coding region, within which several long identical stretches were existed (Figure 5-18 & Figure 5-19). This finding was confirmed by the presence of a strong negative correlation ($r = -0.915$, $P = 0.000079$) between the occurrence of imprecise recombination and the length of the identical sequences (measured by Spearman test Figure 5-32). Based on this, the highest occurrence would be expected to happen at the divergent sites. However, the

highest occurrences were observed within the identical sequence of 2 nts long, and slightly lower when no identical sequence was present (0) i.e. between divergent nucleotides. This maybe a random sampling effect, and could not be confirmed without repeating the experiment and analyse more recombinants.

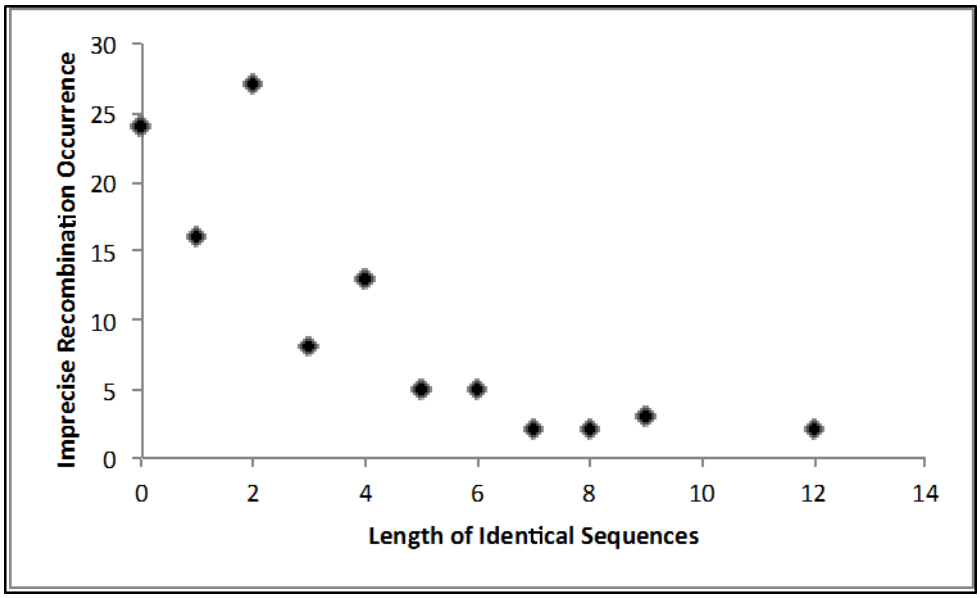


Figure 5-32 Negative correlation between the imprecise recombination occurrence and the length of identical sequence
The x-axis refers to the length of the identical sequences whereas the y-axis reflects the number of imprecise recombination occurrences. The diamonds indicate the relation between the recombination occurrences and the length of the identical sequences.

The tendency of imprecise recombination towards the shorter identical sequences can explain its lower frequency in intratypic in comparison to intertypic recombination, which was found in several studies (Molla et al., 1991, Lowry et al., 2014, Runckel et al., 2013, Tolskaya et al., 1983).

In addition to the presence of long identical sequences within the segment from the end of *VP1*- the middle of *2A*-coding region, a weak negative correlation with the RNA secondary structure was observed in the same region within the imprecise-deletion recombination ($r = -0.365$, $P = 0.019$, see Figure 5-14). Although the correlation was not significant in the insertion but was still noticeable. These findings may imply the presence of a dynamic relation between the RNA secondary structure and sequence identity that can generate a disfavour environment for imprecise recombination (see next chapter).

On the other hand, the presence of the identical sequences did not seem to affect the precise recombination, which was found to occur cross the whole targeted region. However, there did not seem to be a positive correlation between the precise recombination occurrences and the length of the identical sequences (Figure 5-18). Moreover, few precise recombination occurrences were observed at divergent sites between the two viruses. This may suggest that sequence identity is not essentially required for precise recombination to occur; rather it could be favoured by the recombination. Therefore, it is speculated that the role of sequence identity, if it exists, may stem from a need for a sort of alignment to bring the two genomes into the correct position, or for stabilising the structure at a different level and not necessarily adjacent to the junction site. In fact, this may explain the absence of precise recombination and the presence of imprecise occurrences within the *VP1*-coding region part that included in the targeted region (see section 5.7.1 and 5.7.4).

In both precise and imprecise recombination, two types of data were analysed namely flattened and unflattened. The unflattened data, which takes into consideration the number of NGS reads, showed a systematic bias amongst the precise recombination. This was considered a logical reason to refuse to interpret the number of the reads as a reflection of a preferable site (see section 5.7.2). On the contrary, the imprecise did not show such a systematic bias in the unflattened data, therefore, the frequency was considered as an indicator for preferable sites. Based on

the NGS reads, the imprecise-insertion was found to prefer more locations than the imprecise-deletions (Figure 5-11). However, no special correlation was found between the NGS variations and the length of the deleted and inserted fragments or the ability to maintain the open reading frame. Therefore, the consequences of these differences are not clear and open the door for many questions. For example are these molecules evolving in an integral way *i.e.* some of imprecise-deletion out-of-frame recombinants were derived from imprecise-insertion? If this evolving occurs, does it occur in *cis* or *trans*? Furthermore, imprecise recombination flattened data showed that the number of imprecise-deletion recombinants occurred in a higher frequency than expected, which was not the case among the imprecise-insertion recombinants (Figure 5-10). These findings, in addition to the different biases observed within the mononuclutides at the junction site (see Figure 5-25), may suggest that there are different mechanisms that drive the formation of these molecules.

Upon the examination of imprecise recombination, it was observed that the length of deletions and insertions (indels), in the vast majority of recombinants, ranged between 2~10 nts. This is reflected by clustering of recombinant lines at the bottom of the 'length' axis in figure 5-7 B and C. Interestingly, it was noted that more imprecise- recombinants were associated with the dinucleotide identical regions (Figure 5-19). To test for any correlation between these two biases, figure 5-19 was adjusted to show the correlation between the length of deletions and insertions, and the length of the identical regions (Figure 5-33). The relation appeared to be random; the recombinants occurred at identical region of dinucleotide showed variations in the length of deletions and insertions. The lack of correlation was verified by Spearman test, which was carried out on all the imprecise recombinants ($r = 0.185$, $P = 0.05$). This was also applied on imprecise-deletion and -inserion individually (data not shown). Arguably, a possible explanation for observing more recombinants with small insertions would be a technical issue caused by the tendency of PCR to amplify shorter cDNA fragments (see section 6.5). However, this should not be the situation in the case of imprecise-deletion, as the longer the deletion the shorter the recombinant. Therefore, the bias towards the small deletions amongst the imprecise-deletion recombinants was likely caused by the limitations of the PCR assay to detect

any recombinant caused by an RdRp re-association outside of the detectable range on the acceptor strand. Alternatively, it could be a reflection of a random sampling.

Analysing the sequence composition near the junction site showed that RdRp - in the imprecise recombination - favours to switch template immediately after synthesising 'UU' on the donor strand. No difference was observed when tested this on imprecise-deletion and –insertion separately (data not shown). The same finding was observed by others (Nagy and Simon, 1997, King, 1988) and it is believed that the presence of such dinucleotide would promote RdRp slippage. This, in turn, contributes to an accidental incorporation of nontemplated nucleotides at the 3' end of the growing strand. Moreover, it was striking to find that the majority of the imprecise recombination junctions occur next to another dipurine sequences (AG, GG) (see Figure 5-26, D3). In this project, it is not feasible to study how this strong bias towards the dipurine would serve the RdRp to switch template. Further *in vivo* studies are required to place the frequencies of the dinucleotide into an evolutionary context.

In this study, we analysed poliovirus intertypic recombination after 5 hours of the beginning of the infection. The results suggest that recombination at this stage could happen everywhere in the middle of the non-structural region of the viral genome.

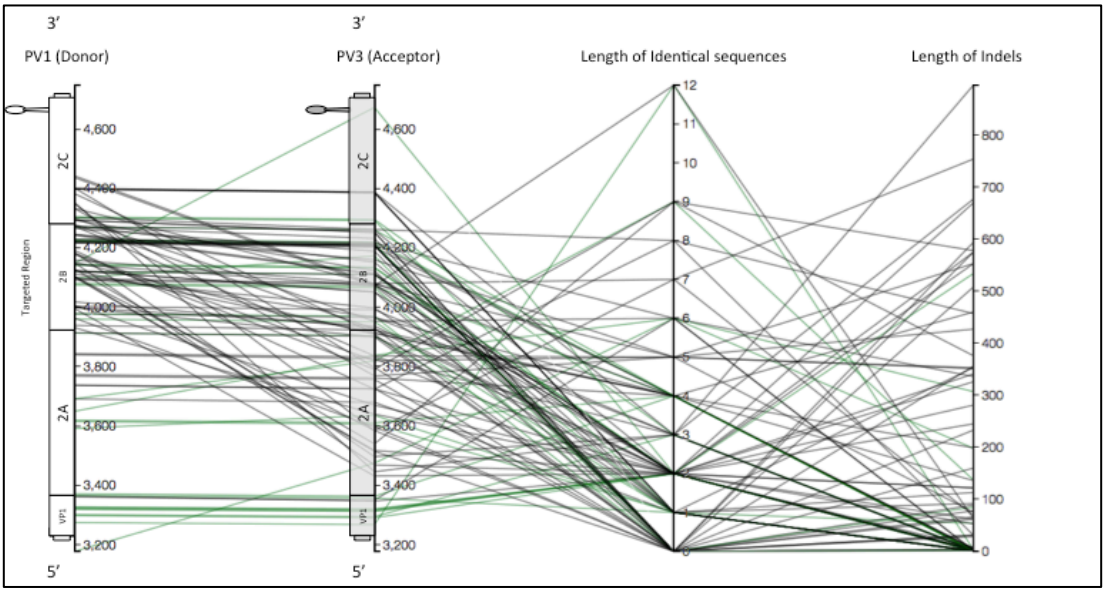


Figure 5-33 Corrolation between the lengths of indels and lenghs of identical sequences
The identified imprecise recombinants and their corresponding length of indels (third axis from the left) were plotted against the length of identical sequences associated with their occurrences (fourth axis from the left). The green lines represent the imprecise-insertion recombinants and the black ones are the imprecise-deletion recombinants.

6 General Discussion

6.1 Current understanding

RNA viruses have gained a reputation as Nature's swiftest evolvers owing to their high mutation rate, high yields, and short replication cycles. Recombination has been shown to play a crucial evolutionary role in the life of RNA viruses. The implications of this are demonstrated by the emergence of modified viruses that have resulted from recombination and are associated with disease outbreaks. For instance, the 2001 outbreak of poliomyelitis in Hispaniola was due to a recombination event that occurred between OPV derived PVs and a non-polio species *C enterovirus* (Kew et al., 2002). Understanding recombination, and how to control it, would help to develop safer vaccines composed of recombination-deficient and genetically stable strains.

This study has benefited from previous recombination studies including the CRE-REP assay, which uses two defective partners that should recombine (at the region between the *VP1* and the CRE) to produce viable progeny (Lowry et al., 2014). This provided an insight into the underlying molecular requirements for recombination by distinguishing between the recombination event itself and the subsequent 'resolution' process. The assay was designed to minimise the competition between the generated recombinant viruses, therefore; reduce the selection to allow for the isolation of a wide range of recombinants regardless of the growth advantages. With this in mind, a more inclusive approach was designed so that in addition to the recombinants with disadvantageous alleles, it can provide an opportunity to characterise the nonviable recombinants. This was achieved by combining a specific RT-PCR amplification with NGS technology to sequence virus progeny. The infection was carried out in a large-scale culture of HeLa cells to increase the absolute number of recombinants to the detectable level (see section 4.3). In contrast to the CRE-REP assay - with which the recombinants were isolated 24-48h post-infection – virus progeny was harvested 5h post-infection even before the virions were produced. Subsequently, whole cell RNA was extracted, cDNA synthesized and PCR amplified. The amplicons were purified and sequenced by Illumina MiSeq. The NGS reads were then analysed using

the ViReMa; a recombination-specialised algorithm that was extensively optimised for the purpose of this study (see section 3.4.3).

Unlike the other proposed methods to study recombination, the method used in this project allows for the isolation of a wider range of recombinants that can occur within the targeted region 5h post infection. For example, viability of recombinants is not a prerequisite *i.e.* defective recombinants, which have lost a fundamental part (required for replication) from their genomes can also be detected. Related to this, recombinants don't have to be in-frame; out-of-frame recombinants are also detectable. Accordingly the non-replicative recombinants, which either have deleterious mutation or are unable to maintain protein-protein compatibility, will also be included. Finally, some recombinants may replicate more efficiently than the others, probably because they have inherited an advantageous mutation. Thus, they can prosper at the expense of other recombinants in the population. The challenge in detecting the low abundant recombinants was overcome by utilising NGS sequencing. The main feature of this approach is that it can focus on recombination at the level of the virus genome, without a subsequent round of selection for viability. For example, if protein-protein compatibility resulted in a recombination junction between X and Y, the CRE-REP assay would only be able to detect those recombinants, which occurred between X and Y. In contrast, utilising NGS on recombinants populations generated at an early stage of infection would allow for the detection of a range of recombination junctions outside of X-Y.

The work presented in this thesis has contributed to the understanding of recombination and its findings are broadly summarised below:

- Precise-recombination appears to be a random event
- Imprecise-recombination appears to be a site-specific event. Detailed analysis of the sequence composition around junction locations did show that imprecise-recombination most likely to occur adjacent to dipurines (AA, AG, GG).
- In-frame and out-of frame imprecise-deletion and imprecise-insertion recombinants– with various deletion/insertion fragments length – can occur early in the infection within the non-structural region P2 (the target region).

- No correlation between the lengths of indel in imprecise recombination and the maintenance of the reading frame.
- The majority of recombinants are precise. At 5h post infection, they appeared twice as much as imprecise recombinants.

6.2 Recombination dynamics and possible recombinants

The essential prerequisite for successful recombination between two viruses is that an individual host (animal or cell) must be infected with both viruses. Taking this into consideration, there are several constraints, which may prevent the recombination process from proceeding. The viruses that participate in forming recombination should pass through all of the stages presented in Figure 6-1 before producing viable recombinants. This section will discuss the assay and the findings of this project within the framework of these constraints

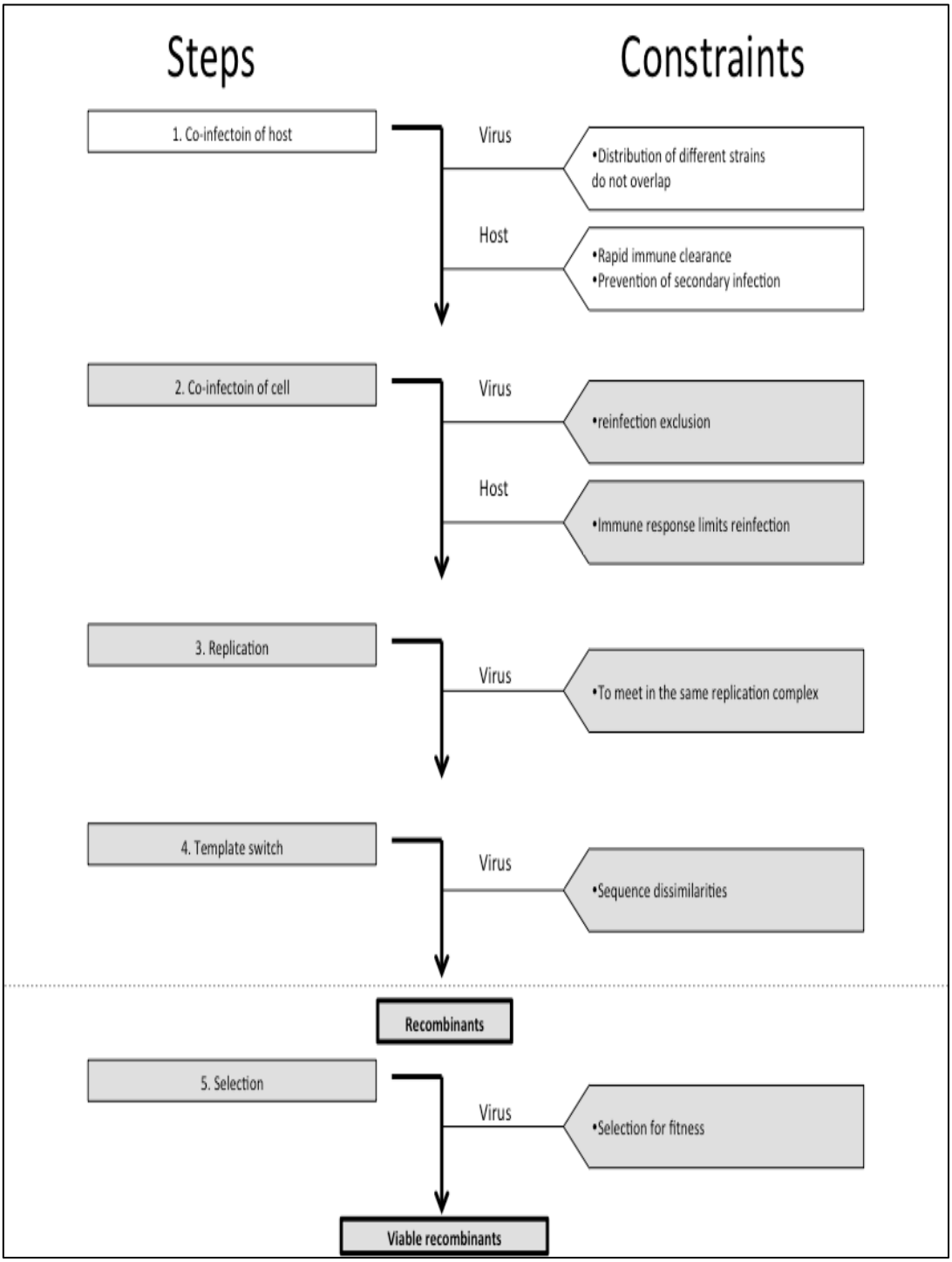


Figure 6-1 A model of the steps required for generating a viable recombinants and possible constraints. It is based on the copy-choice mechanism. Step 1 was not involved in this project; hence the colour of the illustrative boxes is white. This figure is adapted from Worobey M. *et al.* (Worobey and Holmes, 1999)

6.2.1 Dual infection of the same cell

The major constraint viruses can face when entering the host cell is the difficulty in meeting each other at the right place and time (Figure 6-1 step 1). In the case of poliovirus, this restriction can readily be overcome, since recombination events between PV serotypes of OPV are routinely documented in vaccinees (Cammack et al., 1988). The other constraints that might operate at this level are the immune response of the host factors, which can be achieved by either blocking the superinfection or by clearing the virus, hence decreasing the chance of co-infection. Nonetheless, in the studies described here, to ensure the dual infection, this step was bypassed, and instead HeLa cells were infected with two serotypes of poliovirus PV1 and PV3. The infection was carried out at a high multiplicity of infection (10) to make sure that there are enough viruses from both strains to dually infect the vast majority of cells. This, however, does not necessarily imply that every cell will be dually infected with no cells will obtain one virion or none. In fact, this process can be described by Poisson distribution, and expressed by Equation 6-1. The equation can be simplified to calculate the fraction of uninfected cells (k=0), cells with a single infection (k=1), and cells with multiple infections (k>1). Using the equation, the number of cells that receive more than one virus in each flask can be determined. In a culture of $18.6*10^6$ cells the probability of the number of cells that will receive > 1 virus in each flask is p=0.995, which means that ~$18.5*10^6$ cells will receive >1 virions. Clearly, in this large-scale culture, the few cells that do not receive a virus or receive only one virus is negligible, leaving the vast majority of cells with a high chance of receiving more than one virus. As a consequence, the chance for both viruses to meet in one cell is very high.

$$P(K) = \frac{e^{-m}m^k}{k!}$$

Equation 6-1 Poission distribution describing the number of virions each cell recieves
P(k) is the fraction of cells infected by k virus particles, at m MOI. The 'e' reflects the number 2.71828 (Euler's number)

6.2.2 Switching template within replication complex inside the cell

Once PV1 and PV3 are inside the cell, the two viruses have to replicate in the presence of each other for the replicative recombination to proceed. Based on a study by Egger et al., the two viruses would co-localize in a closed-off replication complex near the nuclear periphery, wherein the recombination occurs. In their study, they found a high percentage (85%) of replication complexes containing both viruses early on in the infection. Furthermore, they observed that the first detectable recombinant was 2.5h post infection and was interestingly accompanied by the onset of viral replication (Egger and Bienz, 2002), which may reflect the essential role of recombination for the virus to prosper. Within this locked environment of the replication complex, the two viruses would have an appropriate space to come close to each other, facilitating the RdRp switching template. Accordingly, it seems logical to assume that there are low physical constraints in order for the RdRp to switch templates (Figure 6-1 step 3,4).

Taking Egger et al. into consideration, the two viruses were allowed 2.5 hours inside the replication complexes before being eventually harvested. The results from this work have shown that the RdRp can switch templates at related and unrelated sites in the presence or absence of nucleotide similarities producing all types of recombinants suggested by Lai (Figure 1-3). The analysis of the sequence similarities within all the produced recombinants in this project did not show an apparent influence of sequence similarities on recombination (see section 5.9). This may suggest that recombination occurs via the non-processive model (see section 1.3.4) of the copy-choice mechanism, where the RdRp dissociates from the donor strand but not from the incomplete nascent strand (Jarvis and Kirkegaard, 1991, Jarvis and Kirkegaard, 1992). Notably, the switching template process did not take into account the maintenance of an open reading frame reflected by the presence of out-of-frame and in-frame recombinants in the dataset (Figure 5-7). It seems conceivable that at this stage of infection recombination can occur without a strict selection reflected by the presence of all types of recombinants with various features.

A further investigation including the identical sequences analysis, for both precise and imprecise recombination, was carried out (see section 5.10). Among the imprecise recombination, a negative correlation between long identical sequences

and the occurrence of recombination (see sections 5.10 and 5.12) was found. This was in stark contrast to the precise recombinants, which were found to occur across the whole targeted region regardless of the presence or the length of the identical sequence. This finding was supported by the occurrence of some precise recombination at divergent sites (Figure 5-17 & Figure 5-18). Based on this, it may be possible that the identical sequence *per se* is not essential for precise recombination, and their lengths rather facilitate the process of bringing the two viruses into an aligned position (Romanova et al., 1986). This alignment, by definition, would be a disfavoured environment for the imprecise recombination to occur. Moreover, since the targeted region (nucleotides 3235-4548) is 78% similar between the two viruses, the presence of such an aligning technique (based on the identical sequences) would help the virus to increase the production of the precise recombination at the expense of the imprecise. This, on its turn, can be translated into an evolutionary aspect by ensuring advantageous selection, as precise recombinants would have more chance to replicate (see next section). This is important as it may explain the absence, (or the very low frequency) of the imprecise recombination between two identical viruses (intratypic recombination), even at an early stage and in the absence of the selection pressure (Jarvis and Kirkegaard, 1992).

The question would become then, how recombination could occur at the divergent sites where no such aligning facility is available? It has been hypothesised that RdRp switching may be promoted by a misincorporation event that can lead to stall the enzyme on the RNA template (Dulin et al., 2015). The misincorporation at the 3' end of the nascent strand could anneal with its 'perfect complement' on the acceptor strand and may produce either precise or imprecise recombinants (Agol, 1997). The former will occur if the 'perfect complement' is located in the same position on the donor strand where the misincorporation has taken place on the acceptor stand, while the latter will occur if the nascent strand binds to a distant 'perfect complement' nucleotide. This type of misincorporation cannot be detected in the real dataset, as the misincorporated nucleotide would appear as a correct nucleotide within the recombinant sequence in the aligning process. Therefore, just for the sake of this discussion, this type of misincorporation will be referred to as 'silent misincorporation' (Figure 6-2, left panel, see also section 3.6.1and Read D in Figure

3-11). This is to differentiate it from the other type, where misincorporation may happen without the need for a 'perfect complementation'. This is termed 'non-silent misincorporation' as some of the non-complement nucleotides (Figure 6-2, right panel) would appear as mismatches during the aligning process, and are hence detectable. Unfortunately, this analysis would require additional scripting to extract the relevant information, and due to the time restraints this was not possible. The test can be done by comparing the presence of mutation (non-silent misincorporation) around the junction in the real data and comparing it with a control simulated data.

On the other hand, the analysis of sequence composition around the junction location did show that imprecise recombination favours to occur near dipurines (AA, AG, GG) (Figure 5-26). The significance of this is not clear and can be further investigated by different experiments that involve increasing (or decreasing) the ratio of these dinucleotides in both viruses and measure the outcome. In comparison, precise-recombination did not show any strong evidence for a correlation with either the RNA secondary structure and/or specific sequence motifs (see section 5.11). This was in contrast to other study which proposed that the presence of either an RNA structure and/or sequence motifs (Runckel et al., 2013) was required for recombination to occur. Arguably, although there did not appear to be any correlation with the structure or sequence, the possibility of a combination between structures and/or sequence that serves momentarily to promote RdRp dissociation cannot be ruled out. Alternatively, it could be that the occurrence of precise recombination near identical sequence has masked the presence of specific sequence during the analysis.

To conclude, it seems more conceivable that there is more than one mechanism that drive the recombination process either within the same type of recombination or between different types (precise, imprecise-deletion, and imprecise-insertion recombination).

6.2.3 Competition between recombinants for fitness selection

It is clear that at this early stage of infection all types of recombinants can occur within the targeted region. Whether these recombinants persist, however, depends on the selective separation of the growth advantages and disadvantages among these recombinants (Figure 6-1 step 5). Although recombination can offer growth advantages in some cases (e.g. removing some deleterious mutations, or combining advantageous ones), it can - in a chaotic situation illustrated by the randomness of recombination found in this project - delete more good alleles than it creates. For example, many imprecise-deletion recombinants were detected in the NGS dataset (Figure 5-7 B). As the targeted region (nucleotides 3235-4548) encoded proteins critical for replication such as 2A, 2B, and 2C, any deletion would have a destructive effect on the survival of these molecules. Therefore, it is most likely that these imprecise-deletion molecules would be outcompeted if they were allowed any further inside the cell. i.e. if the virus progeny was harvested at a later stage than 5h post infection. As a consequence, they would be degraded, and not be detected. In fact, this is the major reason for the lack of evidence for the existence of these molecules in recombination studies, and DIs molecules are the only imprecise-deletion recombinants that are published in the literature. Nonetheless, whether some of these defective molecules can restore the missing part of their genomes through a non-replicative recombination (Gmyl et al., 1993) or perhaps replicate at a significantly reduced level (Collis et al., 1992) is unknown. In addition to the imprecise-deletion recombinants, many recombinants (imprecise-insertion recombinants) contained additional sequences (in-frame and out-of-frame) were detected within the targeted regions. The inserted sequence can be either a long insertion that duplicates part of the genome (duplication) or a few nucleotides (see section 1.3.7).

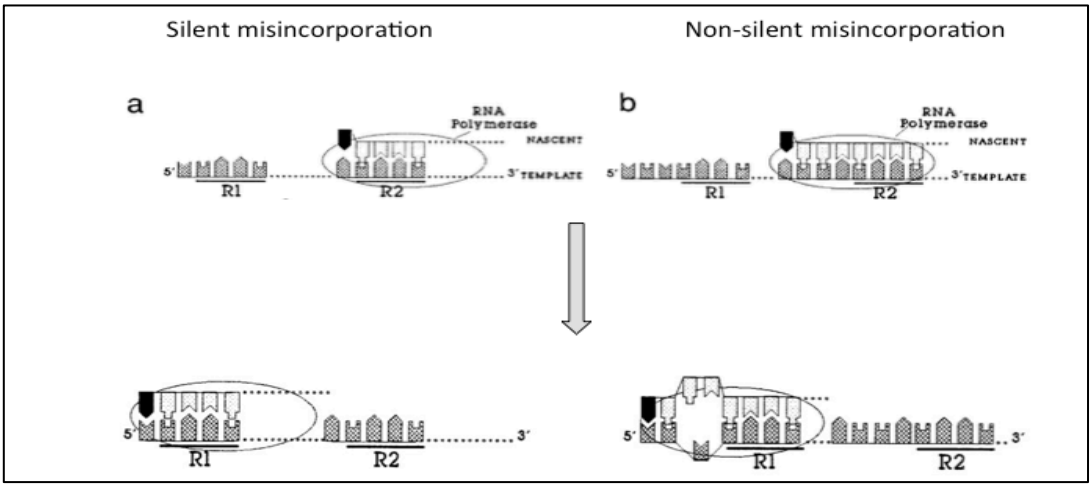


Figure 6-2 A model of RdRp switching recombinants involving a nucleotide misincorporation during the nascent strand synthesis
The complementary between the 3' of the nascent strand and the template is depicted by a key-lock manner. R1 and R2 refer to the association and dissociation sites respectively. The misincorporated nucleotide is represented by a black symbol. A) Silent mutation would happen if the misincorporated nucleotide found a perfect match on the acceptor strand, and this could be either precise or imprecise. In both cases the mutation would not be detectable in the data analysis. B) Non-silent misincorporation would happen if the misincorporated nucleotide couldn't find a perfect match on the acceptor strand. As a result, some nucleotides near the misincorporated one would appear as mismatches in the dataset, and would be therefore detectable.

Unlike the imprecise-deletion recombinants, imprecise-insertion recombinants, as demonstrated by Lowry *et al.*, can replicate, at least the ones that they tested. Although these recombinants are unfit, they can rapidly lose the inserted sequences and regain full fitness upon serial passage (Lowry et al., 2014). Those in CRE-REP assay were isolated in the absence of infective parental genomes, clonally selected and tested separately. However, they would have much lower chance to compete in the presence of infective parental viruses and with such a high MOI of 10 used in this experiment, particularly because they had to go through another step to resolve into the original size by losing the extra sequence. Nevertheless, even if they succeeded to resolve into the original size, and found a safe place to escape the competition with the parental viruses, they might face another constraint, namely protein-protein compatibility. Although these recombinants had a minimal chance to survive in an environment that was limited to a flask and full of competent viruses, it is not clear if this would still be the case in a natural infection environment. Presumably, imprecise-insertion recombinants can travel to a distant site in the same host, or another host, causing a founder effect. In this case, the recombinant can resolve, and its subsequent resolved product can thrive in the absence of the parental genomes.

The resolution of imprecise-insertion recombinants leads to the formation of precise recombinants, which were found to be the predominant recombinants among the results (Figure 5-7). Indeed, this may explain the fact that the frequency of the precise recombinants was found to be double the amount of the imprecise recombinants. This can be attributed to two facts. Firstly, within the allotted time frame of 2.5h (from the beginning of the replication) in the experiment, the recombinants had time to go through replication more than once based on the finding that poliovirus can replicate an average of 6 times before being released (Schulte et al., 2015). Therefore, it seems logical to assume that the obtained precise recombinants may have actually represented a generation that had already gone through the resolution process. Based on this, it can be theorised that the precise recombinants were increasing over time at the expense of the imprecise-insertion recombinants. In fact, this can be mathematically modelled by taking into consideration the generation rate of all recombinants and the resolution rate (Figure 6-3). Simplistically, within a window of 5 hours, precise and imprecise recombinants will be generated, and those with growth advantages can go through a subsequent replication, and the defective ones will accumulate, and at a later stage of the infection they might degrade. This process will continue increasing until the reaction is stopped. Some of these generated recombinants will be imprecise-insertion; therefore, they will go through an additional resolution process, to finally resolve into precise recombinants. Based on the fact that these resolved recombinants might have more chance to replicate (at least by having the same size of the original genome) than other recombinants, they will keep growing constantly. At a certain point (threshold), due to their accumulation and replicative nature, their generation rate will overshadow the overall generation rate of recombinants. The time at which the resolution process would start is yet to be determined; it is logical to assume that this depends on when the imprecise-insertion recombinants would have been created in the population. Additionally, the proportion of these recombinants is an important factor in the process. For example, if imprecise-insertion recombinants appeared at a low frequency in the beginning of the recombinants formation, then it would take more time for the resolved recombinants to establish their own production (Figure 6-3 A). In this case the resolution process would stay part of the overall recombinants generation, and at some point – when enough resolve recombinants produced – the resolved recombinants would establish their own population. The

other scenario would be for the imprecise-insertion recombinants to appear concomitantly with the precise recombinants and at a relatively high frequency (Figure 6-3 B). In this case, the resolution process would establish its own production from the beginning of the recombinants' formation. However, it won't exceed the overall recombinants generation rate, till there are enough resolved recombinants. Finally, the last possible scenario would be for the resolution process to lag behind the overall recombinants generation process. In this situation, the imprecise-insertion recombinants would have been created after some times of the beginning of the recombinants generation (Figure 6-3 C).

Alternatively, the reason for seeing more precise recombinants may simply be attributable to the 'selective advantage'. These recombinants may replicate faster, thus they will appear sooner in the population regardless of the resolution process. This is supported by the finding that ~93% of the total recombination NGS reads belonged to the precise recombinants and only ~7% were derived from the imprecise-recombinants.

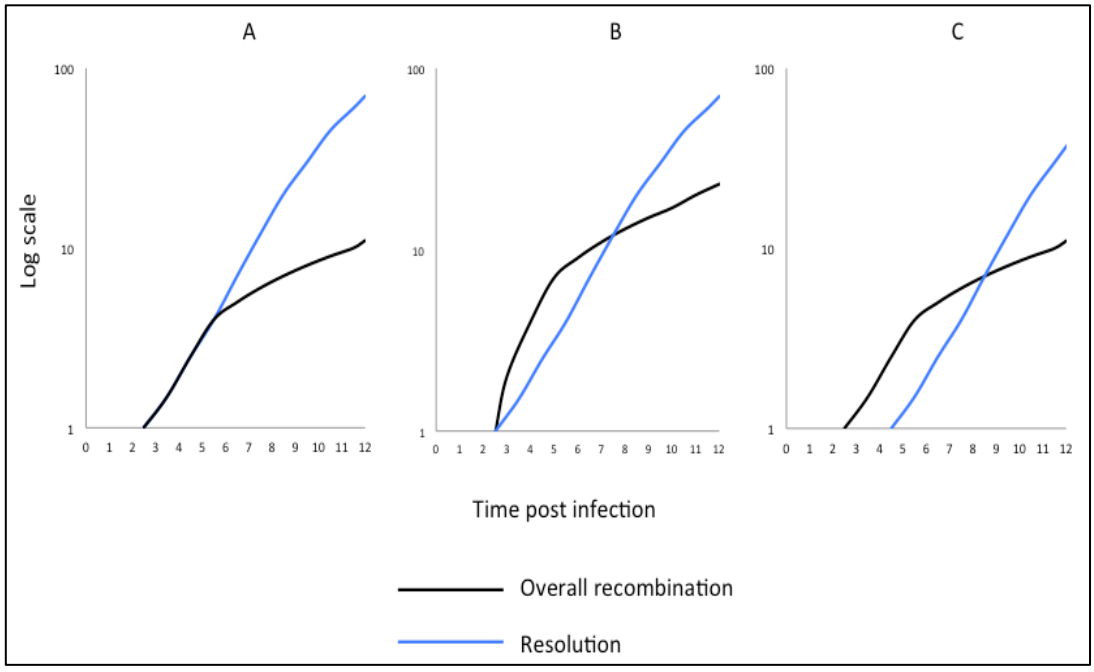


Figure 6-3 Mathematical model representing the outnumbering of precise over the imprecise recombinants
The x-axis represents the time point post infection while the y-axis reflects the possible recombination frequency (log scale). The solid black diagonal line represents the overall recombinants' generation rate. The blue line represents the resolution rate of imprecise-insertion recombinants i.e. the resolved recombinant replication. A) Imprecise-insertion recombinants appeared simultaneously with the recombination onset at a low frequency. B) Imprecise-insertion recombinants appeared simultaneously with the recombination onset at a relatively high frequency C) Imprecise-insertion recombinants appeared later than the initial recombination onset at either high or low frequency.

6.3 Comparison of simulated and experimental data

The bioinformatics pipeline to detect recombination junctions from the NGS dataset was optimised by using simulated datasets. The system, in the presence of 1% mutation within the NGS simulated datasets (Dataset C, Table 3-1), showed a sensitivity of detecting ~64% unique reads from the total number of reads representing recombinants. This was found in a dataset that contained only three different recombinants, and was covered with ~4 million reads. In theory that means that every recombinant has a chance to be covered with 1.3 million reads. By applying the Lander and Waterman equation on this (Equation 3-1) the average coverage for each recombinant was ~300000 reads. This means that every nucleotide had a chance to be seen by 300000 reads, at least once. With such a high coverage, taking into consideration the random nature of the amplicons' fragmentation process, all the recombinants in the dataset would be detectable within the system's sensitivity window of ~64%. Nonetheless, the system would not detect those recombinants that have their locations at the either end of the amplicon. For example, if a recombinant's amplicon in the dataset had its junction location within the first or last 25nts, it would never be picked up by the system (see section 3.7.1 & 3.7.2). Accordingly, the detection of this recombinant is not included within the ~64% sensitivity window.

Analysing the NGS reads from the real data, which was generated under similar conditions to the simulated data (high coverage and long reads), 313 unique recombinants were detected. That means ~64% of the total recombination reads represented 313 unique recombinants. It is possible that the system missed several recombinants because either they did not have the chance to be spanned at the detectable locations by the reads or they contained mutations that prevented the system from recovering them (erroneous reads type 1 and 2 section 3.6.1).

Additionally, there is a small chance that the system reported some inaccurate recombinants caused by a mutation located immediately upstream or downstream to a junction that was situated next to an identical sequence (erroneous reads type 3 section 3.6.1). The reported junction in this case would have shifted from the true

junction by the length of the identical sequence. The fact that there is no way to distinguish these reads in the real dataset – unlike in the simulation dataset – means that few inaccurate recombination junctions (if any) were dealt with as true junctions in the data analysis. Nevertheless, the effect of these reads can be considered negligible based on their minimal chance to appear.

6.4 The reproducibility of the experiment

The experiment developed in this project is a novel approach and has not been applied before to study recombination in the RNA viruses. This means that the results produced by each step were not always predictable and many optimisations were carried out to adjust both the *in vitro* and *in silico* systems. This, besides the large-scale characteristic of the experiments developed in this project, the use of NGS technology, and the large amount of data that could be produced made it difficult in terms of cost and time to repeat the whole experiment. However, the experiment was developed at the University of Warwick where the first successful recombinants' amplification, isolation, and Sanger sequencing were carried out. Subsequently, new samples were produced and sequenced at NIBSC in Dr. Andrew Macadam' laboratory, arranged by Professor David Evans. The experiment was repeated exactly in the same way with the following exceptions; HEp2c cell line (HeLa cells derivatives) was used instead of HeLa cell line, also the virus stock was different from the one used at the University of Warwick. Despite these minor changes The results obtained were similar to the ones obtained at the University of Warwick; a PCR product of the expected size was observed when amplifying recombinants, and no product was observed for the control sample (mixture of RNA viruses) (Figures 4-9 and 5-1). Moreover, the same recombinants found at the University of Warwick by Sanger sequencing were found again in the NGS dataset generated at NIBSC. The only difference was in the clear appearance of the bands that reflected the imprecise recombinants, where only a strong smear appeared at the University of Warwick. This probably can be explained by the fact that different equipment (such as PCR thermocycler, gel electrophoresis, transilluminator) were used, which may had different sensitivity. Nonetheless, the reproducibility could be confirmed by repeating the experiment with three different biological triplicates and compare the results

6.5 Limitations of the experimental and analytical systems

Experiment

The approach of enlarging the scale of infection and using a combination of a specific RT-PCR assay and NGS sequencing has been shown to be successful in amplifying and characterising recombinants that result from wild-type intertypic recombination in poliovirus. However, the cost of this method is high, which imposes some restrictions for applying it more widely, particularly if optimisation steps were needed, which would most probably be the case. Moreover, it is not clear yet if this approach can be used to study recombinants at a late stage of infection. This depends on the time at which all the recombinants would be outcompeted by the parental viruses. This can be measured by carrying out the experiment at diffident time point (see section 6.6.3).

Another considerable drawback of the system is its dependence on the PCR assay, which is known to be biased towards shorter fragments. As a result, recombinants with long insertion may have less chance to be amplified, and hence detected. This could be overcome by using emulsion-PCR rather than the traditional PCR, which would allow uniform amplification of the recombinants' population. This technique was used at an early stage of this project and failed to produce any results when applied on a smaller scale of infection. However, most probably that was a detection sensitivity problem, as the system was not fully optimised by then. The technique could be tried again, and an advantage could be taken from the results produced by this project to properly optimise the emulsion PCR conditions.

Finally, the system is merely qualitative *i.e.* recombinants quantification is not included, which renders the system incapable of measuring the frequency of recombinants as this cannot be inferred from the number of reads (see section 5.3). However, this flaw can be overcome with real-time PCR (see section 6.6.5). Alternatively, a control sample that contains known concentrations of the parental viruses and a spiked-in recombinant can be run in parallel with the rest of the samples. This sample can inform about the relation between the NGS reads and the quantity of the recombinants, which can then be applied on the reset of the samples to estimate the quantity of recombinants. It is worth mentioning that this sample was

included in the initial samples prepared at the university of Warwick. Unfortunately, the construct JC105B was not available at NIBSC, therefore this sample was not included in this project.

ViReMa

The Python implemented algorithm ViReMa with the mode N0X5 was used successfully to detect recombinants from the NGS datasets. However, using this mode would not allow the algorithm to tolerate any mismatch within the 25 nts seed sequence the algorithm extracts from the beginning of the read to initiate the searching process (see section 3.5). As a result, those recombination junctions who were spanned by the ends of the reads could not be detected. To tackle this issue, the reads could be extracted from the unknown file and used as a new dataset for ViReMa to search recombinants within it. This time, the mode N1X5 or N2X5 could be used, as they allow for one or two mutations to happen respectively (see section 3.6), therefore providing more flexibility in finding those reads which were rejected based on the location of the junction. Nonetheless, a care should be taken, as this would decrease the specificity of ViReMa. As a consequence, false positive recombination junctions are possible to be reported. Based on this, one more verification step should be considered. This can be achieved by pulling the reads that correspond to all the reported recombination junctions (see sections 2.8.8 and 2.8.9) and align them against synthetic reference recombinant sequences (see section 3.4.1). Another possibility could be by shortening the seed length and repeat the verification steps.

6.6 Future experiments:

6.6.1 Extend the regions of the genome analysed

Recombinants within the capsid region have rarely been observed, it was suggested that this could be due to the high divergence - at these regions -, the RNA structure, or protein incompatibility (Simmonds and Welch, 2006). Additionally, this region of the genome has been always associated with the occurrence of DIs (Kuge et al., 1986). It would be of interest to apply the approach used in this project to study recombination within the (P1) region. This might help in understanding the mechanisms involved in the generation and evolution of DIs. For example, do imprecise-deletion recombinants occur more than precise or imprecise-insertion? If yes, is the proportion of in-frame higher than the out-of-frame? Is there any correlation between the recombination occurrence and RNA structures or sequence motifs? Studying this region can also be accompanied by studying the P3 region. This can provide a comparable platform through which, the role of some features such as sequence identity and RNA structure can be further investigated. For example the 3D coding region shares a high level of similarities between PV1 and PV3, how this affects the recombination frequency, and how the features of recombinants differ from the P2 and P3 recombinants. This also can provide an insight into the dissociation rate of the RdRp, is it consistent throughout the whole genome?

This can be achieved by designing several specific pairs of primers to target the P1 and P3 regions. Due to the high similarities within the P3 region between both viruses extra care should be taken in designing the P3 primers to maintain specificity. In contrast, designing specific primers for the P1 region should be less challenging. Subsequently, ViReMa can be optimised for these regions. For example, as the 3D coding region (5987-7369nts) contains more sequence identity; there are longer identical stretches here than there are in P2 that can reach 29nts. Based on this, the seed length used in the targeted region in this study (25nts) would not be as specific in this region. Increasing the specificity of the algorithm in this case can be achieved by setting the seed length to 30 or probably more. This should be tested on simulation data, as increasing the length of the seed would enhance the specificity within the 3D coding region, but can also affect the sensitivity. This balance can also

be controlled by the X value, which defines the length of the identical sequence around the junction location. This was set to 5 within the targeted region in this project as the majority of identical sequences ranged between 2 to 5nts in length.

6.6.2 Serial passage of the recombinant population

In a previous study on coronavirus using a specific PCR assay and Sanger sequencing, it was shown that recombinants appeared to be random at an early stage of infection, and after serial passages of the recombinants the recombination sites had clustered in a specific place (Banner and Lai, 1991). Applying this to the recombinant population identified in this study would be a valid experiment in order to study the 'hot-spot' regions. This would contribute to understanding the selection criteria of the fitness recombinants. Moreover, it would provide comparable data with the CRE-REP system, through which a clearer decision could be made if whether the clustering effect observed in CRE-REP was not due to the artificial nature of the assay. For example, would the imprecise-insertion recombinants cluster at *VP1*/*2A* and/or *2A*/*2B* as it was found in CRE-REP? Would the imprecise recombinants disappear altogether? Answering these questions would allow for a wider understanding of the functional constraints that govern the selection of recombinants. This could be achieved by using the supernatant extracted from the four-pooled flasks, to re-infect more HeLa cells. As the sample would contain a large amount of recombinants, it would then be reasonable to carry out the reinfection on a smaller scale e.g. one T175 flask. To make it more inclusive, this could be performed over a time course and in different regions on the genome.

6.6.3 Time course infection (to prove the origin of precise recombinants)

The fact that recombination is a biphasic process (Lowry et al., 2014) which starts to occur simultaneously with the onset of replication (Onodera, 2007) means that, after 5 hours, some recombinants would have been derived from an earlier ones. Therefore, repeating the experiment demonstrated in this study over a time course would be a useful tool to study the evolution of recombinants from the moment they start to appear. Would the vast majority of recombinants be imprecise at 3h post

infection for example? NGS sequencing of the recombinants generated at different time points would shed more light on the factors that govern the ratio of recombinants. This could be achieved by repeating the experiment and harvesting the virus progeny every hour, starting from 2.5h post infection. Although a shorter interval window would be more inclusive, 1h seems more applicable based on the large scale nature of the experiment.

6.6.4 Resolution process (*cis* or *trans*)

Based on the biphasic nature of recombination, the imprecise-insertion recombinants would resolve into precise recombinants with enhanced fitness. The mechanism by which the resolution process occurs is still to be determined. Does it occur in *cis* (part of the insertion is looped-out from the genome) or in *trans* (through another recombination event)? In theory, both mechanisms are possible. The construct JC105B, which was isolated from the CRE-REP assay, could be used as a model in this suggested experiment. It contains an imprecise-insertion recombinant, with 249nts insertion, and was found to go through the resolution process upon serial passage when tested in cell culture (Lowry et al., 2014).

JC105B could be transfected into HeLa cells, and the generated recombinants' population could be sequenced at several passages. However, this would not allow tracing back the origin of the NGS recombination reads and would thus be no way to predict the mechanism by which the original recombinant was generated. To overcome this, a codon-barcode technique (Cho et al., 2015) could be used to insert degenerated nucleotides into JC105B that can serve as unique tags to trace back the molecules. The suggested tags would be unique combinations of synonymous mutations around the insert within JC105B (Figure 6-4). The length of the tag should not be more than 10 to 15nts; the shorter the better. The reason for this is that inserting long tags may increase the effort and the cost of the experiment, besides which it could affect the functionality. The synonymous mutation could be inserted either by gene synthesis (synthesising new construct with the desirable mutation) (Runckel et al., 2013) or by Multiple site-directed mutagenesis (Seyfang and Jin, 2004). Based on this, two JC105B new constructs containing unique tags could be made. Subsequently, these two unique constructs can be co-transfected into HeLa cells, and the virus generations can be monitored at several passages through NGS. A

provisional design for this experiment is illustrated in Figure 6-4. The experimental design involves a PCR amplification that produces a 570nts fragment. By having amplicons at this size, a) no further fragmentation steps are needed and the amplicons can be directly sequenced, b) it will allow the forward and reverse reads to span the junction at sites detectable by ViReMa, and c) it will preserve a suitable length of the amplicon for the sequencings, for example even if the whole insertion has resolved, the remaining fragment (321nts) will still be valid enough to serve as a template for the sequencing with 300nts paired-end reads. The generated reads that spanned the junction location will then be detected by ViReMa, separated into a new FASTQ file and aligned by Bowtie2 with the paired-end aligning mode. For example, as the tag mutations depicted in the Figure 6-4 by x's or o's, if the aligned paired-end reads have x's at their 5' side and o's at their 3' side, this means that the resolution happened in *trans.* The experiment, as described here, is based on two individual constructs; with some experimental modification this could be expanded to include a pool of colonies.

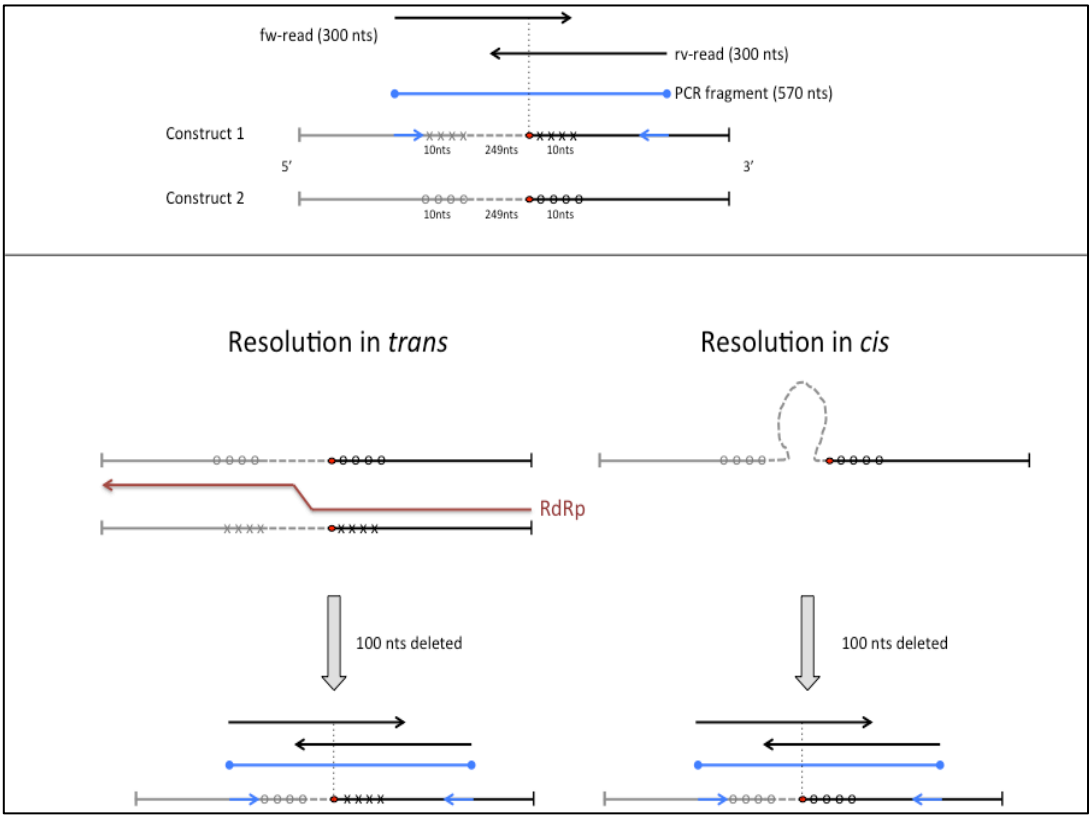


Figure 6-4 NGS provisional experimental design to determine the resolution mechanism
To determine whether the resolution process occurred in *cis* or *trans,* two synthesised JC105B constructs containing unique tags can be used to co-transfect HeLa cells. The top panel describes the specifications of the proposed tests. The constructs are recombinants between PV1 (black line) and PV3 (grey line) that enclose an insertion (grey dashed-line) of 249 nts length. The junction location (red oval) is located downstream to the insertion. The NGS reads (black arrows) are 300 nts length; the small dotted line indicated the sites at which they span the junction location. The locations of the primers are demonstrated by the small blue arrows. The suggested PCR fragment length is 570nts (blue line) to allow the reads to span the junction within sites that are detectable by ViReMa. The two constructs are the same apart from the unique tags they carry at each side of the insertion. These tags are depicted as 'x x x x' in construct 1 and 'o o o o ' in construct 2 and they are 10 nts length. The bottom panel illustrates the two possible mechanisms with a random example of 100 nts being resolved (deleted) from the construct. If resolution occurred in *trans* where the two constructs should recombine to generate a resolved (shorter) recombinant, the generated recombinant would possess a unique tag that has o's on its 5' side and x's on the 3' side. If the resolution occurred in *cis*, however, the resulted molecule would have the same tags as the original constructs.

6.6.5 Real-Time PCR amplification

One of the limitations of this study is the lack of possibilities to calculate the frequency of recombination, which would provide an exact measurement of the proportion of recombinants. To accurately calculate the frequency of recombinants, the recombination products would need to be measured soon after their formation, before any form of selection takes place. This is made possible by repeating the experiment with real-time PCR as described by Jarvis *et al*. (Jarvis and Kirkegaard, 1992). In their study they optimised a condition under which the amount of PCR products was linearly proportional to the amount of input template, which allowed them to quantify recombinants soon after their generation. The same approach can be adapted to modify the current experiment.

6.6.6 Study intra- and interspecies recombinants

Both intra and interspecies recombination have been documented to occur between *enterovirus* members, and in a preliminary study using the CRE-REP assay, recombinants between species *B enterovirus* were successfully recovered. However, it failed to detect recombination between species C and species B (as viable virus progeny). This may have been due to direct genetic incompatibility i.e. the generated recombinants were not viable, or possibly due to the lack of opportunity for both viruses to meet within the same replication complex (Figure 6-1 step 3). These possibilities could be addressed by repeating the experiments to include intra or interspecies viruses.

6.6.7 Study recombination in vivo

The experimental system described in this project provides a platform to study recombination *in vitro* by directly infecting cells with the two viruses. It would be beneficial to apply the method used in this experiment to study recombination within an *in vivo* system. This can be achieved by injecting a solution that contains both viruses directly into the spine of mice (intrathecally). Subsequently, within a short period of time, tissue would be removed from different regions of the mouse including the sites surrounding the injection locations. The samples would then be prepared for NGS and analysed by ViReMa. Unfortunately, this might need to

include more than one mouse in order to increase the absolute number of recombinants. This would allow for the study of recombination in a natural system, which in turn could provide for a broader understanding of the criteria that governs the selection. For example, a correlation between the type of recombination and the location of tissues could be established. Would recombinants that had more chance to travel to a distant tissue have more chance to resolve into precise recombinants?

References

Acevedo, A., Brodsky, L. and Andino, R. (2014) 'Mutational and fitness landscapes of an RNA virus revealed through population sequencing', *Nature,* 505(7485), pp. 686-90.

Agol, V. I. (1997) 'Recombination and Other Genomic Rearrangements in Picornaviruses', *Seminars in VIROLOGY* 8(2), pp. 77-84.

Alexander, L., Lu, H. H. and Wimmer, E. (1994) 'Polioviruses containing picornavirus type 1 and/or type 2 internal ribosomal entry site elements: genetic hybrids and the expression of a foreign gene', *Proc Natl Acad Sci U S A,* 91(4), pp. 1406-10.

Ambros, V., Pettersson, R. F. and Baltimore, D. (1978) 'An enzymatic activity in uninfected cells that cleaves the linkage between poliovirion RNA and the 5' terminal protein', *Cell,* 15(4), pp. 1439-46.

Andino, R., Rieckhof, G. E. and Baltimore, D. (1990) 'A functional ribonucleoprotein complex forms around the 5' end of poliovirus RNA', *Cell,* 63, pp. 369-380.

Bachrach, H. L. and Schwerdt, C. E. (1952) 'Purification Studies on Lansing Poliomyelitis Virus: pH Stability, CNS Extraction and Butanol Purification Experiments'.

Baker, J. C. (1987) 'Bovine viral diarrhea virus: a review', *J Am Vet Med Assoc,* 190(11), pp. 1449-58.

Banner, L. R. and Lai, M. M. (1991) 'Random nature of coronavirus RNA recombination in the absence of selection pressure', *Virology,* 185(1), pp. 441-5.

Bar, K. J., Li, H., Chamberland, A., Tremblay, C., Routy, J. P., Grayson, T., Sun, C., Wang, S., Learn, G. H., Morgan, C. J., Schumacher, J. E., Haynes, B. F., Keele, B. F., Hahn, B. H. and Shaw, G. M. (2010) 'Wide variation in the multiplicity of HIV-1 infection among injection drug users', *J Virol,* 84(12), pp. 6241-7.

Barclay, W., Li, Q., Hutchinson, G., Moon, D., Richardson, A., Percy, N., Almond, J. W. and Evans, D. J. (1998) 'Encapsidation studies of poliovirus subgenomic replicons', *J Gen Virol,* 79 (Pt 7), pp. 1725-34.

Barton, D. J., O'Donnell, B. J. and Flanegan, J. B. (2001) '5' cloverleaf in poliovirus RNA is a cis-acting replication element required for negative-strand synthesis', *Embo j,* 20(6), pp. 1439-48.

Barzon, L., Lavezzo, E., Costanzi, G., Franchin, E., Toppo, S. and Palu, G. (2013) 'Next-generation sequencing technologies in diagnostic virology', *J Clin Virol,* 58(2), pp. 346-50.

Beerenwinkel, N. and Zagordi, O. (2011) 'Ultra-deep sequencing for the analysis of viral populations', *Curr Opin Virol,* 1(5), pp. 413-8.

Belshaw, R., Sanjuan, R. and Pybus, O. G. (2011) 'Viral mutation and substitution: units and levels', *Curr Opin Virol,* 1(5), pp. 430-5.

Bowtie 2: Manual (2016): forge site. Available at: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml - bowtie2-options-score-min. (Accessed: 23/9/2016)

Brister, J. R., Ako-adjei, D., Bao, Y. and Blinkova, O. (2015) 'NCBI Viral Genomes Resource', *Nucleic Acids Res,* 43(Database issue), pp. D571-7.

Brodie, R., Roper, R. L. and Upton, C. (2004) 'JDotter: a Java interface to multiple dotplots generated by dotter', *Bioinformatics,* 20(2), pp. 279-81.

Brown, D. W. (1997) 'Threat to Humans from Virus Infections of Non-human Primates', *Rev Med Virol,* 7(4), pp. 239-246.

Bull, R. A., Luciani, F., McElroy, K., Gaudieri, S., Pham, S. T., Chopra, A., Cameron, B., Maher, L., Dore, G. J., White, P. A. and Lloyd, A. R. (2011) 'Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection', *PLoS Pathog,* 7(9), pp. e1002243.

Cameron, C. E., Oh, H. S. and Moustafa, I. M. (2010) 'Expanding knowledge of P3 proteins in the poliovirus lifecycle', *Future Microbiol,* 5(6), pp. 867-81.

Cammack, N., Phillips, A., Dunn, G., Patel, V. and Minor, P. D. (1988) 'Intertypic genomic rearrangements of poliovirus strains in vaccinees', *Virology,* 167(2), pp. 507-14.

Cao, X. M., Kuhn, R. J. and Wimmer, E. (1993) 'Replication of Poliovirus RNA Containing Two VPg Coding Sequences Leads to a Specific Deletion Event', *Journal of Virology,* 67, pp. 5572-5578.

Cascone, P. J., Carpenter, C. D., Li, X. H. and Simon, A. E. (1990) 'Recombination between satellite RNAs of turnip crinkle virus', *EMBO J,* 9(6), pp. 1709-15.

Chao, L. (1990) 'Fitness of RNA virus decreased by Muller's ratchet', *Nature,* 348(6300), pp. 454-5.

Chao, L. and Tran, T. T. (1997) 'The advantage of sex in the RNA virus phi6', *Genetics,* 147(3), pp. 953-9.

Chetverin, A. B., Chetverina, H. V., Demidenko, A. A. and Ugarov, V. I. (1997) 'Nonhomologous RNA recombination in a cell-free system: evidence for a transesterification mechanism guided by secondary structure', *Cell,* 88(4), pp. 503-13.

Cho, N., Hwang, B., Yoon, J. K., Park, S., Lee, J., Seo, H. N., Huh, S., Chung, J. and Bang, D. (2015) 'De novo assembly and next-generation sequencing to analyse full-length gene variants from codon-barcoded libraries', *Nat Commun,* 6, pp. 8351.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010) 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants', *Nucleic Acids Res,* 38(6), pp. 1767-71.

Coffin, J. M. (1979) 'Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses', *J Gen Virol,* 42(1), pp. 1-26.

Cole, C. N. and Baltimore, D. (1973a) 'Defective interfering particles of poliovirus. II. Nature of the defect', *J Mol Biol,* 76(3), pp. 325-43.

Cole, C. N. and Baltimore, D. (1973b) 'Defective interfering particles of poliovirus. IV. Mechanisms of enrichment', *J Virol,* 12(6), pp. 1414-26.

Cole, C. N., Smoler, D., Wimmer, E. and Baltimore, D. (1971) 'Defective interfering particles of poliovirus: I. Isolation and physical properties', *Journal of Virology,* 7 No. 4, pp. 478-485.

Collis, P. S., O'Donnell, B. J., Barton, D. J., Rogers, J. A. and Flanegan, J. B. (1992) 'Replication of poliovirus RNA and subgenomic RNA transcripts in transfected cells.'.

Crotty, S., Cameron, C. E. and Andino, R. (2001) 'RNA virus error catastrophe: Direct molecular test by using ribavirin'.

Cuervo, N. S., Guillot, S., Romanenkova, N., Combiescu, M., Aubert-Combiescu, A., Seghier, M., Caro, V., Crainic, R. and Delpeyroux, F. (2001) 'Genomic features of intertypic recombinant sabin poliovirus strains excreted by primary vaccinees', *J Virol,* 75(13), pp. 5740-51.

De Jesus, N. H. (2007) 'Epidemics to eradication: the modern history of poliomyelitis', *Virol J,* 4, pp. 70.

Dedepsidis, E., Kyriakopoulou, Z., Pliaka, V. and Markoulatos, P. (2010) 'Correlation between recombination junctions and RNA secondary structure elements in poliovirus Sabin strains', *Virus Genes,* 41(2), pp. 181-91.

Derrien, T., Estelle, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigo, R. and Ribeca, P. (2012) 'Fast computation and applications of genome mappability', *PLoS One,* 7(1), pp. e30377.

Di Giallonardo, F., Zagordi, O., Duport, Y., Leemann, C., Joos, B., Kunzli-Gontarczyk, M., Bruggmann, R., Beerenwinkel, N., Gunthard, H. F. and Metzner, K. J. (2013) 'Next-generation sequencing of HIV-1 RNA genomes: determination of error rates and minimizing artificial recombination', *PLoS One,* 8(9), pp. e74249.

Domingo, E. (1992) 'Genetic variation and quasi-species', *Curr Opin Genet Dev,* 2(1), pp. 61-3.

Domingo, E., Baranowski, E., Escarmís, C., Sobrino, F. and Holland, J. J. (2002) 'Error Frequencies of Picornavirus RNA Polymerases: Evolutionary Implications for Virus Populations'.

Domingo, E. and Holland, J. J. (1997) 'RNA virus mutations and fitness for survival', *Annu Rev Microbiol,* 51, pp. 151-78.

Duffy, S., Shackelton, L. A. and Holmes, E. C. (2008) 'Rates of evolutionary change in viruses: patterns and determinants', *Nat Rev Genet,* 9(4), pp. 267-76.

Dulbecco, R. and Vogt, M. (1954) 'PLAQUE FORMATION AND ISOLATION OF PURE LINES WITH POLIOMYELITIS VIRUSES'.

Egger, D. and Bienz, K. (2002) 'Recombination of poliovirus RNA proceeds in mixed replication complexes originating from distinct replication start sites.', *J Virol,* 76(21), pp. 10960-71.

Egger, D., Teterina, N., Ehrenfeld, E. and Bienz, K. (2000) 'Formation of the Poliovirus Replication Complex Requires Coupled Viral Translation, Vesicle Production, and Viral RNA Synthesis'.

Elena, S. F., Carrasco, P., Daros, J. A. and Sanjuan, R. (2006) 'Mechanisms of genetic robustness in RNA viruses', *EMBO Rep,* 7(2), pp. 168-73.

Enders, J. F., Weller, T. H. and Robbins, F. C. (1949) 'Cultivation of the Lansing Strain of Poliomyelitis Virus in Cultures of Various Human Embryonic Tissues', *Science,* 109(2822), pp. 85-7.

Evans, D. M., Dunn, G., Minor, P. D., Schild, G. C., Cann, A. J., Stanway, G., Almond, J. W., Currey, K. and Maizel, J. V., Jr. (1985) 'Increased neurovirulence associated with a single nucleotide change in a noncoding region of the Sabin type 3 poliovaccine genome', *Nature,* 314(6011), pp. 548-50.

Fawcett, T. (2006) 'An introduction to ROC analysis', *Pattern Recognition Letters,* 27(8), pp. 861-874.

Flexner, S. and Amoss, H. L. (1919) 'PERSISTENCE OF THE VIRUS OF POLIOMYELITIS IN THE NASOPHARYNX', *J Exp Med,* 29(4), pp. 379-95.

Flint, J., Racaniello, V. R., Rall, G. F., Skalka, A. M. and Enquist, L. W. (2015) *Principles of Virology, Fourth Edition.* 4TH edn.: ASM press, p. 121-123.

Forss, S. and Schaller, H. (1982) 'A tandem repeat gene in a picornavirus', *Nucleic Acids Res,* 10(20), pp. 6441-50.

Freistadt, M. S., Vaccaro, J. A. and Eberle, K. E. (2007a) 'Biochemical characterization of the fidelity of poliovirus RNA-dependent RNA polymerase', *Virol J,* 4, pp. 44.

Freistadt, M. S., Vaccaro, J. A. and Eberle, K. E. (2007b) 'Biochemical characterization of the fidelity of poliovirus RNA-dependent RNA polymerase', *Virology Journal,* 4(1), pp. 1.

Frey, K. G., Herrera-Galeano, J. E., Redden, C. L., Luu, T. V., Servetas, S. L., Mateczun, A. J., Mokashi, V. P. and Bishop-Lilly, K. A. (2014) 'Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood', *BMC Genomics,* 15(1), pp. 96.

Ge, H., Liu, K., Juan, T., Fang, F., Newman, M. and Hoeck, W. (2011) 'FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution'.

Gmyl, A. P., Korshenko, S. A., Belousov, E. V., Khitrina, E. V. and Agol, V. I. (2003) 'Nonreplicative homologous RNA recombination: promiscuous joining of RNA pieces?', *Rna,* 9(10), pp. 1221-31.

Gmyl, A. P., Pilipenko, E. V., Maslova, S. V., Belov, G. A. and Agol, V. I. (1993) 'Functional and genetic plasticities of the poliovirus genome: quasi-infectious RNAs modified in the 5'-untranslated region yield a variety of pseudorevertants', *J Virol,* 67(10), pp. 6309-16.

Goodfellow, I., Chaudhry, Y., Richardson, A., Meredith, J., Almond, J. W., Barclay, W. and Evans, D. J. (2000) 'Identification of a cis-Acting Replication Element within the Poliovirus Coding Region', *J Virol: Vol. 10*, pp. 4590-600.

Grada, A. and Weinbrecht, K. (2013) 'Next-Generation Sequencing: Methodology and Application', *Journal of Investigative Dermatology,* 133(8).

Guest, S., Pilipenko, E., Sharma, K., Chumakov, K. and Roos, R. P. (2004) 'Molecular mechanisms of attenuation of the Sabin strain of poliovirus type 3', *J Virol,* 78(20), pp. 11097-107.

Harismendy, O. and Frazer, K. (2009) 'Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology', *Biotechniques,* 46(3), pp. 229-31.

Heath, L., van der Walt, E., Varsani, A. and Martin, D. P. (2006) 'Recombination patterns in aphthoviruses mirror those found in other picornaviruses', *J Virol,* 80(23), pp. 11827-32.

Henn, M. R., Boutwell, C. L., Charlebois, P., Lennon, N. J., Power, K. A., Macalalad, A. R., Berlin, A. M., Malboeuf, C. M., Ryan, E. M., Gnerre, S., Zody, M. C., Erlich, R. L., Green, L. M., Berical, A., Wang, Y., Casali, M., Streeck, H., Bloom, A. K., Dudek, T., Tully, D., Newman, R., Axten, K. L., Gladden, A. D., Battis, L., Kemper, M., Zeng, Q., Shea, T. P., Gujja, S., Zedlack, C., Gasser, O., Brander, C., Hess, C., Gunthard, H. F., Brumme, Z. L., Brumme, C. J., Bazner, S., Rychert, J., Tinsley, J. P., Mayer, K. H., Rosenberg, E., Pereyra, F., Levin, J. Z., Young, S. K., Jessen, H., Altfeld, M., Birren, B. W., Walker, B. D. and Allen, T. M. (2012) 'Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection', *PLoS Pathog,* 8(3), pp. e1002529.

Herold, J. and Andino, R. (2001) 'Poliovirus RNA replication requires genome circularization through a protein-protein bridge', *Mol Cell,* 7(3), pp. 581-91.

Hogle, J. M., Chow, M. and Filman, D. J. (1985) 'Three-dimensional structure of poliovirus at 2.9 A resolution', *Science,* 229(4720), pp. 1358-65.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S. and VandePol, S. (1982) 'Rapid evolution of RNA genomes', *Science,* 215(4540), pp. 1577-85.

Holmes, E. C. (2003) 'Error thresholds and the constraints to RNA virus evolution', *Trends Microbiol,* 11(12), pp. 543-6.

Hurst, L. D. and Peck, J. R. (1996) 'Recent advances in understanding of the evolution and maintenance of sex', *Trends Ecol Evol,* 11(2), pp. 46-52.

Hyypia, T., Hovi, T., Knowles, N. J. and Stanway, G. (1997) 'Classification of enteroviruses based on molecular and biological properties', *J Gen Virol,* 78 (Pt 1), pp. 1-11.

Illumina (2015a) *Nextera DNA Library Prep Kit*. Available at: http://www.illumina.com/products/nextera_dna_library_prep_kit.html (Accessed: 23/11/2015).

Illumina (2015b) *Paired-End Sequencing*. Available at: http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html (Accessed: 23/11/2015).

Innan, H. and Kondrashov, F. (2010) 'The evolution of gene duplications: classifying and distinguishing between models', *Nat Rev Genet,* 11(2), pp. 97-108.

Invitrogen (2015) *SuperScript® III Reverse Transcriptase*. Available at: http://www.bioscience-events.com/leipzig/Nickson-Invitrogen-qPCR-Leipzig.pdf (Accessed: 3/11/2015.

Jang, S. K., Davies, M. V., Kaufman, R. J. and Wimmer, E. (1989) 'Initiation of protein synthesis by internal entry of ribosomes into the 5' nontranslated region of Encephalomyocarditis Virus RNA in vivo', *Journal of Virology,* 63 (4), pp. 1651-1660.

Jarvis, T. C. and Kirkegaard, K. (1991) 'The polymerase in its labyrinth: mechanisms and implications of RNA recombination', *Trends Genet,* 7(6), pp. 186-91.

Jarvis, T. C. and Kirkegaard, K. (1992) 'Poliovirus RNA recombination: mechanistic studies in the absence of selection', *EMBO J,* 11(8), pp. 3135-45.

Kaplan, G. and Racaniello, V. R. (1988) 'Construction and characterization of poliovirus subgenomic replicons', *J Virol,* 62(5), pp. 1687-96.

Kew, O., Morris-Glasgow, V., Landaverde, M., Burns, C., Shaw, J., Garib, Z., André, J., Blackman, E., Freeman, C. J., Jorba, J., Sutter, R., Tambini, G., Venczel, L., Pedreira, C., Laender, F., Shimizu, H., Yoneyama, T., Miyamura, T., van Der Avoort, H., Oberste, M. S., Kilpatrick, D., Cochi, S., Pallansch, M. and de Quadros, C. (2002) 'Outbreak of poliomyelitis in Hispaniola associated with circulating type 1 vaccine-derived poliovirus', *Science,* 296(5566), pp. 356-9.

Kew, O. M., Sutter, R. W., de Gourville, E. M., Dowdle, W. R. and Pallansch, M. A. (2005) 'Vaccine-derived polioviruses and the endgame strategy for global polio eradication', *Annu Rev Microbiol,* 59, pp. 587-635.

Khatchikian, D., Orlich, M. and Rott, R. (1989) 'Increased viral pathogenicity after insertion of a 28S ribosomal RNA sequence into the haemagglutinin gene of an influenza virus', *Nature,* 340(6229), pp. 156-7.

Kim, D. and Salzberg, S. L. (2011) 'TopHat-Fusion: an algorithm for discovery of novel fusion transcripts', *Genome Biol,* 12(8), pp. R72.

King, A. M. (1988) 'Preferred sites of recombination in poliovirus RNA: an analysis of 40 intertypic cross-over sequences', *Nucleic Acids Res,* 16(24), pp. 11705-23.

King, A. M., McCahon, D., Slade, W. R. and Newman, J. W. (1982) 'Recombination in RNA', *Cell,* 29(3), pp. 921-8.

Kirkegaard, K. and Baltimore, D. (1986) 'The mechanism of RNA recombination in poliovirus', *Cell,* 47(3), pp. 433-43.

Kitamura, N., Semler, B. L., Rothberg, P. G., Larsen, G. R., Adler, C. J., Dorner, A. J., Emini, E. A., Hanecak, R., Lee, J. L., Van der Werf, S., Anderson, C. W. and Wimmer, E. (1981) 'Primary structure, gene organisation and polypeptide expression of poliovirus RNA', *Nature,* 291, pp. 547-553.

Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M. and Seelow, D. (2011) 'Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing', *PLoS One,* 6(11), pp. e28240.

Kuge, S., Kawamura, N. and Nomoto, A. (1989) 'Strong inclination toward transition mutation in nucleotide substitutions by poliovirus replicase', *J Mol Biol,* 207(1), pp. 175-82.

Kuge, S., Saito, I. and Nomoto, A. (1986) 'Primary structure of poliovirus defective-interfering particle genomes and possible generation mechanisms of the particles', *J Mol Biol,* 192(3), pp. 473-87.

Lai, M. M. (1992) 'RNA recombination in animal and plant viruses', *Microbiol Rev,* 56(1), pp. 61-79.

Lander, E. S. and Waterman, M. S. (1988) 'Genomic mapping by fingerprinting random clones: a mathematical analysis', *Genomics,* 2(3), pp. 231-9.

Lander, E. S. W., Michael S (2015) *Genomic mapping by fingerprinting random clones: A mathematical analysis.*

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods,* 9, pp. 357-359.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome Biol,* 10(3), pp. R25.

Lauring, A. S. and Andino, R. (2010) 'Quasispecies theory and the behavior of RNA viruses', *PLoS Pathog,* 6(7), pp. e1001005.

Ledinko, N. (1963) 'Genetic recombination with poliovirus type 1: studies of crosses between a normal horse serum-resistant mutant and several guanidine-resistant mutants of the same strain', *Virology,* 20(1), pp. 107-119.

Lee, Y., Nomoto, A., Detjen, B. M. and Wimmer, E. (1977) 'A protein covalently linked to poliovirus genome RNA', *Proceedings of the National Academy of Sciences of the United States of America,* 74, pp. 59-63.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, G. P. D. P. (2009) 'The Sequence Alignment/Map format and SAMtools'.

Lowry, K., Woodman, A., Cook, J. and Evans, D. J. (2014) 'Recombination in enteroviruses is a biphasic replicative process involving the generation of greater-than genome length 'imprecise' intermediates', *PLoS Pathog,* 10(6), pp. e1004191.

Lundquist, R. E., Sullivan, M. and Maizel, J. V., Jr. (1979) 'Characterization of a new isolate of poliovirus defective interfering particles', *Cell,* 18(3), pp. 759-69.

Luscombe, N. M., Greenbaum, D. and Gerstein, M. (2001) 'What is bioinformatics? A proposed definition and overview of the field', *Methods Inf Med,* 40(4), pp. 346-58.

Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. and Chinnaiyan, A. M. (2009) 'Transcriptome sequencing to detect gene fusions in cancer', *Nature,* 458(7234), pp. 97-101.

Makino, S., Keck, J. G., Stohlman, S. A. and Lai, M. M. (1986) 'High-frequency RNA recombination of murine coronaviruses.'.

Malim, M. H. and Emerman, M. (2001) 'HIV-1 sequence variation: drift, shift, and attenuation', *Cell,* 104(4), pp. 469-72.

McClure, M. A., Holland, J. J. and Perrault, J. (1980) 'Generation of defective interfering particles in picornaviruses', *Virology,* 100(2), pp. 408-18.

McInerney, P., Adams, P. and Hadi, M. Z. (2014) 'Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase', *Mol Biol Int,* 2014, pp. 287430.

Medvedev, P., Fiume, M., Dzamba, M., Smith, T. and Brudno, M. (2010) 'Detecting copy number variation with mated short reads', *Genome Res,* 20(11), pp. 1613-22.

Mendelsohn, C. L., Wimmer, E. and Racaniello, V. R. (1989) 'Cellular receptor for poliovirus: molecular cloning, nucleotide sequence, and expression of a new member of the immunoglobulin superfamily', *Cell,* 56(5), pp. 855-65.

Metzker, M. L. (2010) 'Sequencing technologies - the next generation', *Nat Rev Genet,* 11(1), pp. 31-46.

Meyers, G., Tautz, N., Dubovi, E. J. and Thiel, H. J. (1991) 'Viral cytopathogenicity correlated with integration of ubiquitin-coding sequences', *Virology,* 180(2), pp. 602-16.

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., Shaw, P. D. and Marshall, D. (2013) 'Using Tablet for visual exploration of second-generation sequencing data'.

Mitra, R. D. and Church, G. M. (1999) 'In situ localized amplification and contact replication of many individual DNA molecules', *Nucleic Acids Res,* 27(24), pp. e34.

Miura, F., Uematsu, C., Sakaki, Y. and Ito, T. (2005) 'A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3′-end subsequences'.

Molla, A., Paul, A. V. and Wimmer, E. (1991) 'Cell-free, de novo synthesis of poliovirus', *Science,* 254(5038), pp. 1647-51.

Muller, H. J. (1964) 'THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE', *Mutat Res,* 106, pp. 2-9.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. (1986) 'Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction', *Cold Spring Harb Symp Quant Biol,* 51 Pt 1, pp. 263-73.

Murray, M. G., Bradley, J., Yang, X. F., Wimmer, E., Moss, E. G. and Racaniello, V. R. (1988a) 'Poliovirus host range is determined by a short amino acid sequence in neutralization antigenic site I', *Science,* 241, pp. 213-215.

Murray, M. G., Kuhn, R. J., Arita, M., Kawamura, N., Nomoto, A. and Wimmer, E. (1988b) 'Poliovirus type 1/type 3 antigenic hybrid virus constructed in vitro elicits type 1 and type 3 neutralizing antibodies in rabbits and monkeys'.

Nagy, P. D. and Bujarski, J. J. (1993) 'Targeting the site of RNA-RNA recombination in brome mosaic virus with antisense sequences'.

Nagy, P. D. and Bujarski, J. J. (1995) 'Efficient system of homologous RNA recombination in brome mosaic virus: sequence and structure requirements and accuracy of crossovers', *J Virol,* 69(1), pp. 131-40.

Nagy, P. D. and Bujarski, J. J. (1997) 'Engineering of homologous recombination hotspots with AU-rich sequences in brome mosaic virus.'.

Nagy, P. D. and Simon, A. E. (1997) 'New insights into the mechanisms of RNA recombination', *Virology,* 235(1), pp. 1-9.

Nayak, D. P. (1980) 'Defective interfering influenza viruses', *Annu Rev Microbiol,* 34, pp. 619-44.

NCBI (2016) *FASTA format*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/.

Nora, T., Charpentier, C., Tenaillon, O., Hoede, C., Clavel, F. and Hance, A. J. (2007) 'Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment', *J Virol,* 81(14), pp. 7620-8.

Oh, H. S., Pathak, H. B., Goodfellow, I. G., Arnold, J. J. and Cameron, C. E. (2009) 'Insight into poliovirus genome replication and encapsidation obtained from studies of 3B-3C cleavage site mutants', *J Virol,* 83(18), pp. 9370-87.

Olson, N. H., Kolatkar, P. R., Oliveira, M. A., Cheng, R. H., Greve, J. M., McClelland, A., Baker, T. S. and Rossmann, M. G. (1993) 'Structure of a human rhinovirus complexed with its receptor molecule', *Proc Natl Acad Sci U S A,* 90(2), pp. 507-11.

Omata, T., Kohara, M., Kuge, S., Komatsu, T., Abe, S., Semler, B. L., Kameda, A., Itoh, H., Arita, M., Wimmer, E. and Nomoto, A. (1986) 'Genetic analysis of the attenuation phenotype of poliovirus type 1', *Journal of Virology,* 58, pp. 348-358.

Onodera, K. (2007) 'Selection for 3'-end triplets for polymerase chain reaction primers', *Methods Mol Biol,* 402, pp. 61-74.

Parameswaran, P., Charlebois, P., Tellez, Y., Nunez, A., Ryan, E. M., Malboeuf, C. M., Levin, J. Z., Lennon, N. J., Balmaseda, A., Harris, E. and Henn, M. R. (2012) 'Genome-wide patterns of intrahuman dengue virus diversity reveal associations with viral phylogenetic clade and interhost diversity', *J Virol,* 86(16), pp. 8546-58.

Park, S. H., Goo, J. M. and Jo, C. H. (2004) 'Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists', *Korean J Radiol,* 5(1), pp. 11-8.

Parnell, L. D., Lindenbaum, P., Shameer, K., Dall'Olio, G. M., Swan, D. C., Jensen, L. J., Cockell, S. J., Pedersen, B. S., Mangan, M. E., Miller, C. A. and Albert, I. (2011) 'BioStar: an online question & answer resource for the bioinformatics community', *PLoS Comput Biol,* 7(10), pp. e1002216.

Pathak, K. B. and Nagy, P. D. (2009) 'Defective Interfering RNAs: Foes of Viruses and Friends of Virologists', *Viruses,* 1(3), pp. 895-919.

Paul, A. V., Boom, J. H. v., Filippov, D. and Wimmer, E. (1998) 'Protein-primed RNA synthesis by purified poliovirus RNA polymerase', *Nature,* 393(6682), pp. 280-284.

Pearson, W. R. (2013) 'An Introduction to Sequence Similarity ("Homology") Searching', *Curr Protoc Bioinformatics,* 0 3.

Percy, N., Barclay, W. S., Sullivan, M. and Almond, J. W. (1992) 'A poliovirus replicon containing the chloramphenicol acetyltransferase gene can be used to study the replication and encapsidation of poliovirus RNA', *J Virol,* 66(8), pp. 5040-6.

Perrault, J. (1981) 'Origin and replication of defective interfering particles', *Curr Top Microbiol Immunol,* 93, pp. 151-207.

Picornavirus Home (2016). Available at: http://www.picornaviridae.com/ 30 Apr 2016).

Pilipenko, E. V., Gmyl, A. P. and Agol, V. I. (1995) 'A model for rearrangements in RNA genomes', *Nucleic Acids Res,* 23(11), pp. 1870-5.

Pilipenko, E. V., Maslova, S. V., Sinyakov, A. N. and Agol, V. I. (1992) 'Towards identification of cis-acting elements involved in the replication of enterovirus and rhinovirus RNAs: a proposal for the existence of tRNA-like terminal structures'.

PMC, E. (1984) 'Experimental production of fatal mucosal disease in cattle. - Abstract - Europe PMC'.

Pringle, C. R. (1965) 'EVIDENCE OF GENETIC RECOMBINATION IN FOOT-AND-MOUTH DISEASE VIRUS', *Virology,* 25, pp. 48-54.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics,* 26(6), pp. 841-2.

Racaniello, V. (2007) *Picornaviridae: The viruses and their replication.* 4th edn.: Lippincott Williams & Wilkins, Philadelphia, PA.

Racaniello, V. R. and Baltimore, D. (1981) 'Cloned poliovirus complementary DNA is infectious in mammalian cells', *Science,* 214(4523), pp. 916-9.

Radford, A. D., Chapman, D., Dixon, L., Chantrey, J., Darby, A. C. and Hall, N. (2012) 'Application of next-generation sequencing technologies in virology', *J Gen Virol,* 93(Pt 9), pp. 1853-68.

Reichhardt, T. (1999) 'It's sink or swim as a tidal wave of data approaches', *Nature,* 399(6736), pp. 517-20.

Rice, W. R. (2002) 'Experimental tests of the adaptive significance of sexual recombination', *Nature Reviews Genetics,* 3(4), pp. 241-251.

Richter, D. C., Ott, F., Auch, A. F., Schmid, R. and Huson, D. H. (2008) 'MetaSim: a sequencing simulator for genomics and metagenomics', *PLoS One,* 3(10), pp. e3373.

ROBBINS, F. C. and ENDERS, J. F. (1950) 'Tissue Culture Techniques in the Study of Animal Viruses.'.

Romanova, L. I., Blinov, V. M., Tolskaya, E. A., Viktorova, E. G., Kolesnikova, M. S., Guseva, E. A. and Agol, V. I. (1986) 'The primary structure of crossover regions of intertypic poliovirus recombinants: a model of recombination between RNA genomes.', *Virology,* 155(1), pp. 202-13.

Routh, A. and Johnson, J. E. (2013) 'Discovery of functional genomic motifs in viruses with ViReMa–a Virus Recombination Mapper–for analysis of next-generation sequencing data', *Nucleic acids research*, pp. gkt916.

Routh, A. and Johnson, J. E. (2014) 'Discovery of functional genomic motifs in viruses with ViReMa-a Virus Recombination Mapper-for analysis of next-generation sequencing data', *Nucleic Acids Res,* 42(2), pp. e11.

Routh, A., Ordoukhanian, P. and Johnson, J. E. (2012) 'Nucleotide-resolution profiling of RNA recombination in the encapsidated genome of a eukaryotic RNA virus by next-generation sequencing', *J Mol Biol,* 424(5), pp. 257-69.

Runckel, C., Westesson, O., Andino, R. and Derisi, J. L. (2013) 'Identification and manipulation of the molecular determinants influencing poliovirus recombination', *PLoS Pathog,* 9(2), pp. e1003164.

Rust, R. C., Landmann, L., Gosert, R., Tang, B. L., Hong, W., Hauri, H. P., Egger, D. and Bienz, K. (2001) 'Cellular COPII proteins are involved in production of the vesicles that form the poliovirus replication complex', *J Virol,* 75(20), pp. 9808-18.

Sambrook, J., Russell, D. W., Maniatis, T. and Fritsch, E. F. (2000) *Molecular Cloning: A Laboratory Manual.* New York: Cold Spring Harbor Laboratory Press, p. 999.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proc Natl Acad Sci U S A,* 74(12), pp. 5463-7.

Sanjuán, R. (2012) 'From Molecular Genetics to Phylodynamics: Evolutionary Relevance of Mutation Rates Across Viruses', *PLoS Pathog: Vol. 5*.

Scheel, T. K., Galli, A., Li, Y. P., Mikkelsen, L. S., Gottwein, J. M. and Bukh, J. (2013) 'Productive homologous and non-homologous recombination of hepatitis C virus in cell culture', *PLoS Pathog,* 9(3), pp. e1003228.

Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S. M., Grocock, R., Henderson, S., Khrebtukova, I., Kingsbury, Z., Luo, S., McBride, D., Murray, L., Menju, T., Timbs, A., Ross, M., Taylor, J. and Bentley, D. (2012) 'Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns', *Blood,* 120(20), pp. 4191-6.

Schulte, M. B., Draghi, J. A., Plotkin, J. B., Andino, R. and Goffs, S. P. (2015) 'Experimentally guided models reveal replication principles that shape the mutation distribution of RNA viruses'.

Schwartz, S. L. and Farman, M. L. (2010) 'Systematic overrepresentation of DNA termini and underrepresentation of subterminal regions among sequencing templates prepared from hydrodynamically sheared linear DNA molecules', *BMC Genomics,* 11, pp. 87.

Seyfang, A. and Jin, J. H. (2004) 'Multiple site-directed mutagenesis of more than 10 sites simultaneously and in a single round', *Anal Biochem,* 324(2), pp. 285-91.

Shendure, J. and Ji, H. (2008) 'Next-generation DNA sequencing', *Nat Biotechnol,* 26(10), pp. 1135-45.

Simmonds, P. and Welch, J. (2006) 'Frequency and dynamics of recombination within different species of human enteroviruses.', *J Virol,* 80(1), pp. 483-93.

Simon-Loriere, E. and Holmes, E. C. (2011) 'Why do RNA viruses recombine?', *Nature Reviews Microbiology,* 9(8), pp. 617-626.

Simon-Loriere, E. and Holmes, E. C. (2013) 'Gene duplication is infrequent in the recent evolutionary history of RNA viruses', *Mol Biol Evol,* 30(6), pp. 1263-9.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A. and Ponting, C. P. (2014) 'Sequencing depth and coverage: key considerations in genomic analyses', *Nature Reviews Genetics,* 15, pp. 121-132.

Stothard, P. (2000) 'The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences', *Biotechniques,* 28(6), pp. 1102, 1104.

Svitkin, Y. V., Imataka, H., Khaleghpour, K., Kahvejian, A., Liebig, H. D. and Sonenberg, N. (2001) 'Poly(A)-binding protein interaction with elF4G stimulates picornavirus IRES-dependent translation', *Rna,* 7(12), pp. 1743-52.

Syed, F., Grunenwald, H. and Caruccio, N. (2009) 'Optimized library preparation method for next-generation sequencing', *Nature Methods,* 6(10).

Tapparel, C., Cordey, S., Junier, T., Farinelli, L., Van Belle, S., Soccal, P. M., Aubert, J. D., Zdobnov, E. and Kaiser, L. (2011) 'Rhinovirus genome variation during chronic upper and lower respiratory tract infections', *PLoS One,* 6(6), pp. e21163.

Tolskaya, E. A., Romanova, L. A., Kolesnikova, M. S. and Agol, V. I. (1983) 'Intertypic recombination in poliovirus: genetic and biochemical studies', *Virology,* 124(1), pp. 121-32.

Tolskaya, E. A., Romanova, L. I., Blinov, V. M., Viktorova, E. G., Sinyakov, A. N., Kolesnikova, M. S. and Agol, V. I. (1987) 'Studies on the recombination between RNA genomes of poliovirus: the primary structure and nonrandom distribution of crossover regions in the genomes of intertypic poliovirus recombinants', *Virology,* 161(1), pp. 54-61.

Toyoda, H., Franco, D., Fujita, K., Paul, A. V. and Wimmer, E. (2007) 'Replication of poliovirus requires binding of the poly(rC) binding protein to the cloverleaf as well as to the adjacent C-rich spacer sequence between the cloverleaf and the internal ribosomal entry site', *J Virol,* 81(18), pp. 10017-28.

Toyoda, H., Kohara, M., Kataoka, Y., Suganuma, T., Omata, T., Imura, N. and Nomoto, A. (1984) 'Complete nucleotide sequences of all three poliovirus serotype genomes. Implication for genetic relationship, gene function and antigenic determinants', *J Mol Biol,* 174(4), pp. 561-85.

Toyoda, H., Nicklin, M. J., Murray, M. G., Anderson, C. W., Dunn, J. J., Studier, F. W. and Wimmer, E. (1986) 'A second virus-encoded proteinase involved in proteolytic processing of poliovirus polyprotein', *Cell,* 45(5), pp. 761-70.

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) 'TopHat: discovering splice junctions with RNA-Seq', *Bioinformatics,* 25(9), pp. 1105-11.

Töpfer, A., Zagordi, O., Prabhakaran, S., Roth, V., Halperin, E. and Beerenwinkel, N. (2013) 'Probabilistic Inference of Viral Quasispecies Subject to Recombination', *J Comput Biol: Vol. 2*, pp. 113-23.

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. and Andino, R. (2006) 'Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population', *Nature,* 439(7074), pp. 344-8.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F. and Liu, J. (2010) 'MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery'.

Wells, V. R., Plotch, S. J. and DeStefano, J. J. (2001) 'Determination of the mutation rate of poliovirus RNA-dependent RNA polymerase', *Virus Res,* 74(1-2), pp. 119-32.

Worobey, M. and Holmes, E. C. (1999) 'Evolutionary aspects of recombination in RNA viruses', *J Gen Virol,* 80 (Pt 10), pp. 2535-43.

Xie, C. and Tammi, M. T. (2009) 'CNV-seq, a new method to detect copy number variation using high-throughput sequencing', *BMC Bioinformatics,* 10, pp. 80.

Zhang, Y., Wang, H., Zhu, S., Li, Y., Song, L., Liu, Y., Liu, G., Nishimura, Y., Chen, L., Yan, D., Wang, D., An, H., Shimizu, H., Xu, A. and Xu, W. (2010) 'Characterization of a rare natural intertypic type 2/type 3 penta-recombinant vaccine-derived poliovirus isolated from a child with acute flaccid paralysis.', *J Gen Virol,* 91(Pt 2), pp. 421-9.

Appendix